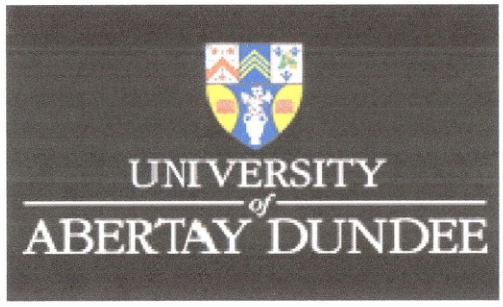


A phylogenetic study of Ubiquitin Conjugating Enzymes and structural analysis of human UBE2J proteins

Joy Mukherjee

A Thesis submitted in partial fulfilment of the University of Abertay Dundee for the Degree of Doctor of Philosophy

School of Contemporary Sciences
University of Abertay Dundee

October 2006

ABSTRACT

The ubiquitin proteasome system (UPS) is responsible for the degradation and turnover of proteins in eukaryotes. As such it is a key process that is involved in normal and in some cases, abnormal cellular functions. Ubiquitin conjugating enzymes (UBCs) are key components of the UPS and may serve as therapeutic targets.

The aims of this project are the structural and functional analyses of UBCs in eukaryotic organisms whose genomes have been fully sequenced, and also the review of the nomenclature of yeast and human UBCs. The main findings were:

1) The successful construction of Phylogenetic trees containing all 14 yeast UBCs and UBC-like proteins and their homologues in selected species whose genomes have been fully sequenced. The phylogenetic tree consists of 15 different branches. Thirteen of the branches contain a member of the yeast UBC or UBC-like family (MMS2) and their homologues. The remaining two branches contain firstly the human UBE2L proteins and secondly the TSG101 UBC like proteins, both of which branches do not appear to have any yeast orthologues.

2) The UBC phylogenetic study was also used to identify the previously unknown UBC active site of the *Drosophila* transcription factor, $TAF_{II}250$. Using multiple sequence manual alignments of known $TAF_{II}250$ and UBC protein sequences a putative UBC active site in *Drosophila melanogaster* and *Apis mellifera* $TAF_{II}250$ was successfully identified, and was shown to share approximately 70% homology to the known UBC PROSITE signature.

3) Using multiple sequence alignments the hitherto unknown PROSITE signature of the clinically important UBE2J family was identified. This PROSITE signature is very different from all other UBCs suggesting that this family of enzymes has significant structural changes at their active sites. Homology modelling proved to be a successful approach to obtain structural information of the UBE2J1 active site. Superimposition studies using the previously solved structures for human UBE2J2 and human UBC9 were carried out. Significant differences were observed near the active sites of human UBE2J proteins compared to the active site for human UBC9.

DECLARATION

I, Joy Mukherjee, hereby certify that this thesis has been written by me, that it is a record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

Date..12:.02:.2007 Signature of Candidate..[redacted]

I hereby certify that the candidate has fulfilled the conditions of the resolution and regulations appropriate for the degree of Doctor of Philosophy at the University of Abertay, Dundee and that the candidate is qualified to submit this thesis in application for the degree.

Date...12.02.2007 Signature of Director of studies[redacted]

In submitting this thesis to the University of Abertay, Dundee I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby.

Date..12:.02:.2007 Signature of candidate..[redacted]

ACKNOWLEDGEMENTS

I would like to thank the ESF funding body for funding the first two years of my University tuition fees.

I would like to thank my supervisors, Dr. Douglas Lester and Dr Eldridge Buultjens, for sharing with me their enthusiasm for the research, and providing me with opportunities, encouragement, and support throughout my PhD, both with my work and in developing my own ideas and interests. Although the work that I had carried out in my research was not in any collaboration, it benefited greatly from the excellent help and assistance freely given by Professor Ron Hay (University of St. Andrews) and Dr. Dimitris Xirodimus (Post Doctorate under the supervision of Prof. Ron Hay) at the BMS building St Andrews University where I had carried out the crystallography experiments. I greatfully acknowledge their help and assistance throughout the time of research at the St. Andrews University. I would additionally like to thanks Dr Whang Ting and Dr. Steven for their kind support at the St. Andrews University.

I would also like to extend my gratitude to Dr. Sharon Kelly (University of Glasgow) and Dr. Jane Potter (of St. Andrews University) for helping me to do the CD spectrum analysis, which was carried out at the University of St. Andrews.

Dr. Frank Wright of the SCRI (Scottish Crop Research Institute) had guided me in the phylogenetic analysis, which helped me a lot to get the analysis done. I would like to thank him especially for that.

I would also like to extend my gratitude to my mother and second brother, for their support during the period of my research.

I would like to especially thank my parents in law who have treated me as their own son, and have supported me (both mentally and financially) during this period of my research.

Last but not the least I would like to say that without the strong support of my wife, I would not have been able to reach up to this stage. Tanima (my Wife) has gone through the struggle during this period along with me, and have given me all possible support that was required.

CONTENTS PAGE

CONTENTS PAGE

CONTENTS PAGE

CONTENTS PAGE

FIGURES **PAGES**

FIGURES **PAGES**

FIGURES **PAGES**

TABLES **PAGES**

TABLES **PAGES**

ABBREVIATIONS, SYMBOLS AND NOTATIONS

ATP	Adenosine Triphosphate
APL	acute promyelocytic leukemia
BLOSUM	Block Substitution Matrix
BLAST	Basic local alignment search tool
CDC	Cell division cycle
CFTR	Cystic Fibrosis Transmembrane Conductance Regulator
CPY	carboxypeptidase Y
CD	circular dichroism
DNA	Deoxy-Ribonucleic acid
EBI	European Bioinformatics Institute
ERAD	Endoplasmic Reticulum Associated Degradation
ERQC	Endoplasmic Reticulum quality control
EMBOSS	The European Molecular Biology Open Software Suite.
E2	It is the abbreviation of UBE2. All yeast UBC orthologues have been given a nomenclature naming starting with UBE2 followed by the alphabet corresponding to each individual UBCs. So at times UBC and E2 have been used together.
FBPase	Fructose-1,6-bisphosphatase
HAT	Histone Acetyl Transferase
HMGCoA	3-hydroxy-3-methyl-glutaryl-CoA
HMM	Hidden Markov Models
KDa	Kilo Dalton
MHC	Major histocompatibility complex
MSA	Multiple sequence alignment
MPD	2-methyl 2,4-pentadiol
NCBI	National Center for Biotechnology Information
NTK	N-Terminal Kinase
NEDD8	neural precursor cell expressed developmentally downregulated gene 8
NMR	Nuclear magnetic resonance
NCoR	nuclear receptor corepressor
PML	promyelocytic leukaemia
PAM	point accepted mutation
PCR	Polymerase Chain Reaction

pI	Isoelectric point
pH	"p" stands for "potenz" (this means the potential to be) and the "H" stands for Hydrogen
PEGs	Polyethylene glycols
PDB	Protein data bank
RUB1	related to ubiquitin 1 protein
RB	Retinoblastomaprotein
RAR-α	retinoic acid receptor-α
SUMO	Small Ubiquitin like modifier
SMRT	silencing mediator of retinoic acid and thyroid hormone receptor
SDS-PAGE	Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
TAF	TBP Associated Factor
TBP	TATA Box-binding Protein
TFIID	Transcription Factor complex
TIGR	The institute for genome research.
TSG101	Tumor suppressor gene 101
UPS	Ubiquitin Proteasome System
UBC	Ubiquitin Conjugating Enzymes
UBL	Ubiquitin like
UGGT	UDP-glucose glycoprotein:glucosyltransferase

Organism Latin name	**abbreviation**	**name commonly used in this thesis**
Anopheles gambiae	Ag	mosquito
Arabidopsis thaliana	At	same as the Latin name
Caenorhabditis elegans	Ce	same as the Latin name
Drosophila melanogaster	Dm	same as the Latin name
Homo sapiens	Hs	human
Mus musculus	Mm	mouse
Neurospora crassa	Nc	same as the Latin name
Oryza sativa	Os	rice
Plasmodium falciparum	Pf	same as the Latin name
Plasmodium yoelli yoelli	Pyy	same as the Latin name
Schizosaccharomyces pombe	Sp	pombe
Saccharomyces cerevisiae	Sc	yeast

The amino acids, symbols, and codons

Amino acids	Symbols		Codons
Alanine	Ala	A	GCA, GCC, GCG, GCU
Aspartic acid	Asp	D	GAC, GAU
Asparagine	Asn	N	AAC, AAU
Arginine	Arg	R	AGA, AGG, CGA, CGC, CGG, CGU
Cysteine	Cys	C	UGC, UGU
Glutamine	Gln	Q	CAA, CAG
Glycine	Gly	G	GGA, GGC, GGG, GGU
Glutamic acid	Glu	E	GAA, GAG
Histidine	His	H	CAC, CAU
Isoleucine	Ile	I	AUA, AUC, AUU
Lysine	Lys	K	AAA, AAG
Leucine	Leu	L	UUA, UUG, CUA, CUC, CUG, CUU
Methionine	Met	M	AUG
Proline	Pro	P	CCA, CCC, CCG, CCU
Phenylalanine	Phe	F	UUC, UUU
Serine	Ser	S	AGC, AGU, UCA, UCC, UCG, UCU
Threonine	Thr	T	ACA, ACC, ACG, ACU
Tryptophan	Trp	W	UGG
Tyrosine	Tyr	Y	UAC UAU
Valine	Val	V	GUA, GUC, GUG, GUU

CHAPTER ONE

GENERAL INTRODUCTION TO UBIQUITIN PROTEASOME SYSTEM (UPS)

1.1 The importance of the UPS in eukaryotes

Proteins destined for degradation by the proteasome are tagged by the covalent attachment of one or more molecules of the protein ubiquitin. The proteasome only recognizes ubiquitinated proteins and subsequently digests them into small peptides (see Figure 1.1). This Ubiquitin Proteasome System (UPS) is important in regulating the half-lives of proteins that have critical regulatory functions within the cell, such as those controlling gene transcription and cell cycle progression, including transcription factors, tumor suppressors such as p53, oncoproteins, certain cell surface receptors, cyclins and related proteins (Bence, Sampat and Kopito, 2001). The UPS is also responsible for the degradation of mutant or misfolded proteins and has been implicated in the processing of MHC (major histocompatibility complex) restricted class I antigens. Additionally the ubiquitin–proteasome pathway is responsible for antigen presentation and the degradation of mutant or misfolded proteins that are processed by the endoplasmic reticulum (ER) by a process known as Endoplasmic Reticulum Associated Degradation or ERAD (Kostova and Wolf, 2003; Plemper and Wolf, 1999; Bonifacino and Weissman, 1998; Lester et al., 2000).

Ubiquitin is a heat stable small 76 amino acid protein, which is apparently available in all eukaryotes. It is not, however found in eubacteria or archaebacteria. Ubiquitin is found in the cellular compartments such as the nucleus, the cytosol, and the cell membrane surface. It is either free or covalently attached to other proteins and all of its related functions are mediated through its linkage to cellular proteins. The post-translational conjugation of ubiquitin to proteins usually leads to branched ubiquitin-protein conjugates. This conjugate has the carboxyl C-terminus of the ubiquitin protein covalently attached to the ε-amino group of a lysine residue in the target protein (Jentsch, 1992).

1.2 Enzymes involved in the UPS

Ubiquitination is a multistep process and it involves four classes of enzymes (see Figure 1.1). In the first step ubiquitin is first activated by an E1 or ubiquitin activating enzyme. This reaction requires ATP hydrolysis.

In the second step the activated ubiquitin is transferred to a conserved cysteine residue on one of the several ubiquitin conjugating enzymes (UBCs) or E2s (E2 is the

abbreviation of UBE2. All yeast UBC orthologues have been given a nomenclature naming starting with UBE2 followed by the alphabet corresponding to each individual UBCs. UBC and E2 have been used together at times). The third step involves participation of a third class of enzymes E3 (ubiquitin ligase), which catalyses the formation of an isopeptide linkage between the C-terminal glycine of the ubiquitin moiety and the ε-amino group of an internal lysine in the target protein. A polyubiquitin chain is synthesised by the successive transfer of ubiquitin molecules to lysine 29, or lysine 48 of the previous ubiquitin moiety. In contrast protein modification by lysine 63 linked chains or by a single ubiquitin moiety (mono-ubiquitination), seem to trigger other functions like DNA repair, gene expression, and protein sorting (see Figure 1.3) (Plemper and Wolf, 1999; Zhang et al., 2003). In the fourth and final step, the ubiquitin chain is elongated by an ubiquitin-chain elongating factor (E4), which assembles the branched polyubiquitin chain (Hoppe, 2005; Hershko and Ciechanover, 1992; Hamilton et al., 2001; Ciechanover, 1994; Richly et al., 2005; Chau et al., 1989).

1.3 Degradation of polyubiquitinated proteins by the 26S proteasome

Ubiquitinated proteins are finally unfolded and degraded by the 26S proteasome (see Figure 1.1). The 26S proteasome is a complex of 20S core and the 19S cap or regulatory particles. The 20S core is composed of four stacked rings, each containing seven different α or β subunits with an overall α7β7β7α7 geometry. The function of the 19S cap is to recognize the ubiquitinated protein, bind and unfold it, and also to regulate the opening of the 20S core. After the degradation of the protein, the ubiquitin is set free where it is again recycled as shown in the figure below, to carry out another set of ubiquitination reactions and the cycle is repeated. This is represented by the Figure 1.1 (Kostova and Wolf, 2003).

Figure 1.1 The ubiquitin proteasome system

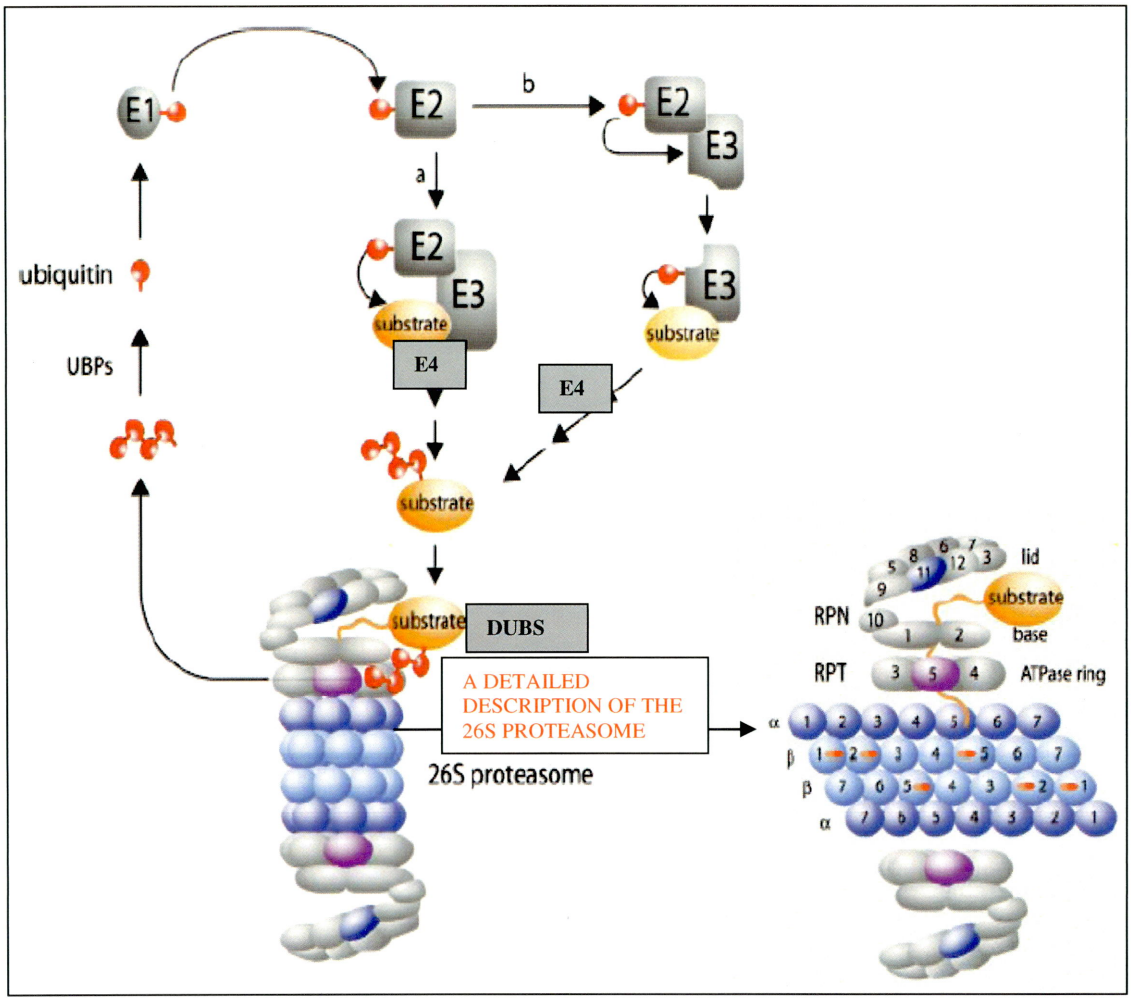


Figure 1.1 shows the ubiquitin proteasome system. This involves four enzymes, where in the first step E1 called the ubiquitin activating enzyme forms a thioester bond with the C terminal glycine residue of ubiquitin. This activated ubiquitin is then transferred to the active site cysteine residue of the ubiquitin conjugating enzyme residue E2. E3 a ubiquitin ligase mediating the attachment of the ubiquitin to the amino group of the lysine residue of substrate proteins through an isopeptide bond. The fourth and the final step is the chain elongation step by the involvement of the enzyme E4. This is then recognized and degraded by the 26S proteasome. After the degradation of the protein that was destined to be degraded the polyubiquitin chain is set free by deubiquinating enzymes (DUBs) (Amerik et al., 2004; Soboleva et al., 2004; Guterman and Glickman, 2004) where it is again recycled, into monoubiquitin by (UBPs) enzymes as shown in the figure above. These monoubiquitin molecules are then free to carry out another set of ubiquitination reactions (Kostova and Wolf, 2003).

1.4 Mono-ubiquitination and multi-ubiquitination

Most proteins targeted for degradation by the UPS are polyubiquitinated by the reaction described in Figure 1.1. However some proteins, e.g. histones, which are not targeted for degradation, can be modified by monoubiquitination, resulting in a change in structure and function of these proteins. Monoubiquitination of protein requires the involvement of only the E1 and E2 enzyme activity. Activation of gene expression may be correlated with monoubiquitination of histones, however the functional connections between histone ubiquitination and activation of gene expression is currently unknown, but presumably it may involve alteration of chromatin structure allowing transcription factors access to the exposed DNA (Pham and Sauer, 2000).

It has been found that most histones (e.g. histone H1), are monoubiquitinated by conventional UBC E2s but not by E3s (Haas, Bright and Jackson, 1988). An unusual example of monoubiquitination activity by the *Drosophila* transcription factor $TAF_{II}250$, was demonstrated by Pham and Sauer (2000).

Research was carried out by Pham and Sauer (2000) to understand the role of histone ubiquitination in transcriptional regulation. As a first step towards this understanding they had tried to identify enzymes that ubiquitinate histones in *Drosophila* embryonic nuclear extract using an activity gel assay (Acetyltransferase Activity Gel Assay- The crude macronuclear extracts or column purified fractions were analysed for histone acetyltransferase activity following electrophoresis in SDS or polyacrylamide gels containing calf thymus histones or bovine serum albumin). To purify the identified activity from the nuclear extract, gel filtration chromatography was carried out, and it was observed that the fractionation pattern from the activity gel assay resembled the pattern described for the general transcription factor TAFIID (Pham and Sauer, 2000; Brownell and Allis, 1995).

The TFIID transcription initiation complex is composed of TBP and multiple TAFs. It nucleates the assembly of RNA polymerase II and other initiation factors (TFIIA, TFIIB, TFIIE, TFIIF, TFIIH) at the core promoter of protein coding genes. The *Drosophila melanogaster* TFIID is composed of TBP and eight other TAFs namely TAF250, TAF150, TAF110, TAF80, TAF60, TAF40, TAF30α and TAF30β. TAF250 is the largest subunit of TAFIID, and is required for progression through G1/S of the cell cycle and repression of apoptosis. $TAF_{II}250$ is required for regulating the cell cycle, cell differentiation, cell proliferation and cell survival. $TAF_{II}250$ consists of two

independent protein kinase domains and a HAT (histone acetyltransferase) domain (Wassarman et al., 2000; Jacobson et al., 2000). The binding of activator proteins to the enhancer region of the target genes initiates the transcriptional activation. Within the chromatin, the chromosomal DNA associates with histones (H1, H2A, H2B, H3, H4) to form nucleosomes. This can inhibit the interaction of transcriptional factors and the general transcriptional machinery with target genes, which activates transcription (Pham and Sauer, 2000; Ruppert, Wang and Tjian, 1993; Hisatake et al., 1993).

As the largest TAFII subunit is TAFII250 within the TFIID complex, the precipitated activity may correspond to *Drosophila* $TAF_{II}250$. Antibodies to ubiquitin detected ubiquitin at a position that coincides with $dTAF_{II}250$, suggesting that $dTAF_{II}250$ ubiquinates histone H1. The nuclear membrane assay and the molecular weight of known enzymes in the ubiquitin pathway suggest that the nuclear membrane bound $dTAF_{II}250$ most likely does not interact with the E1, E2, or E3 enzymes. As it is known that monoubiquitination involves at least E1 and E2 activities, the result implies that $dTAF_{II}250$ may have intrinsic E1 and E2 activities (Jacobson et al., 2000; Mizzen and Allis, 2000).

It was detected in the membrane assay that ubiquitin/H1 conjugates resisted reducing agents, which suggests that $dTAF_{II}250$ may mediate a covalent bond between ubiquitin and H1 by means of isopeptide linkages which indicates that $dTAF_{II}250$ may have intrinsic E2 activity, since this reaction is characteristic for E2 enzymes.

Generally E1 enzymes require ATP to conjugate with ubiquitin to form the thioester bond. Hence to find out whether $dTAF_{II}250$ has the E1 activity, Pham and Sauer (2000) investigated the capability of $dTAF_{II}250$ to conjugate with ubiquitin by means of thioester bond. It was found that $dTAF_{II}250$ conjugated with ubiquitin in an ATP dependent manner in the absence of a reducing agent. This suggests that $dTAF_{II}250$ may have both E1 and E2 activities and may therefore be a "ubac" (ubiquitin activating and conjugating). The E1/E2 activities for $dTAF_{II}250$ are known, but the PROSITE signature of the E2 domain is unknown. The current view of the functional domains of $dTAF_{II}250$ is shown in Figure 1.2.

Figure 1.2

The functional domains of dTAF$_{II}$250

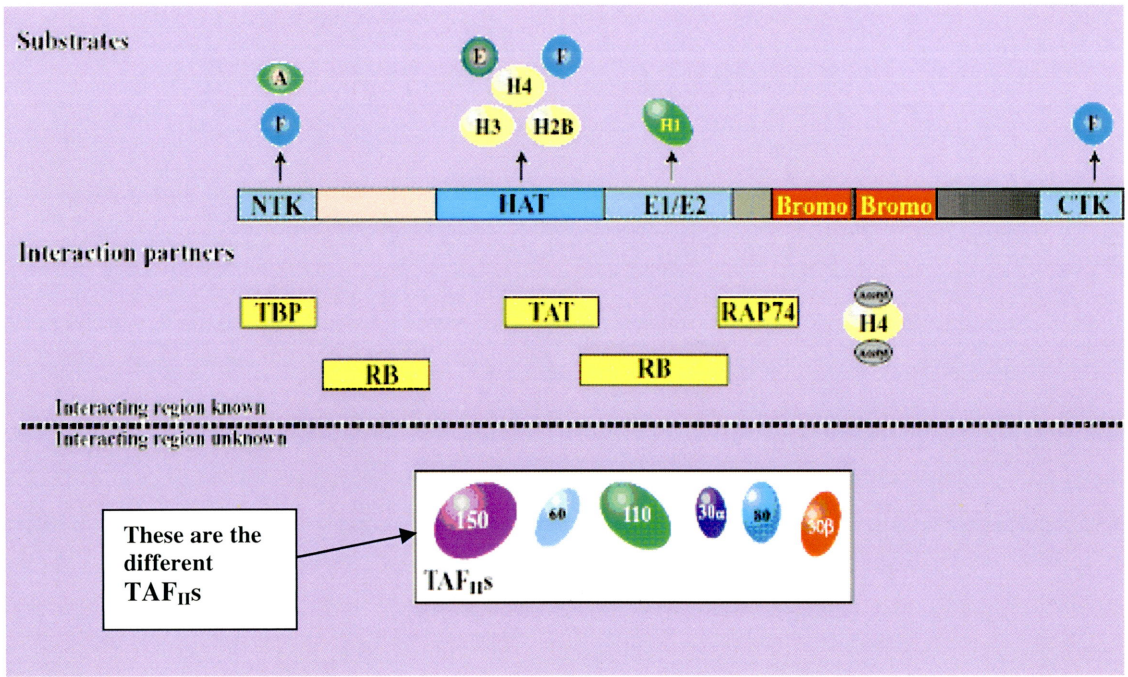


Figure 1.2 is a diagram of metazoan TAF$_{II}$250. NTK is the N-terminal kinase domain, CTK is the C-terminal kinase domain, HAT is the histone acetyl transferase domain, E1/E2 are the ubiquitin activating/conjugating domain respectively, and Bromo is the Bromodomain that are indicated in the linear TAF$_{II}$250 protein. The interacting proteins are indicted below the TAF$_{II}$250 protein. TFIIA(A), TFIIE(E), TFIIF(F), TATA-binding protein(TBP), retinoblastomaprotein (RB), TFIIFα (RAP75) are the substrates and interacting partners (Wassarman and Sauer, 2001)

As can be seen from Figure 1.1 and Figure 1.3, ubiquitin itself is often a substrate for further ubiquitination, which results in the formation of a polyubiquitin chain. There are seven lysine residues in ubiquitin and these residues are used in the different polyubiquitin chain formation signalling different functions. Proteins modified by the lysine 48 (K48) or lysine 29 (K29) linked chains, are degraded by the proteasome; whereas those modified by the lysine 63 (K63) linked chains or by the single ubiquitin moiety (or monoubiquitination) have other functions such as DNA repair, gene expression, protein sorting as shown in the Figure 1.3.

Figure 1.3 Ubiquitin modifications and cellular response

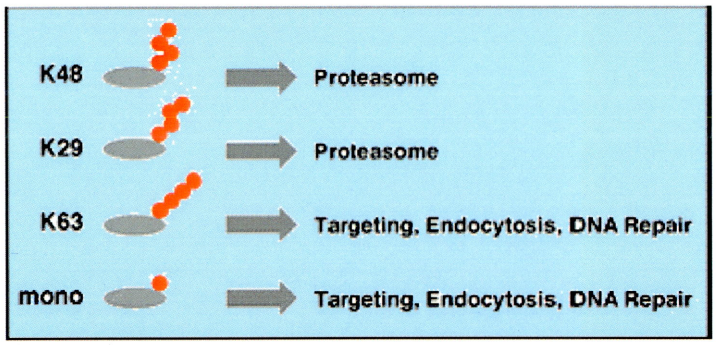


Figure 1.3 illustrates the alteration of functions by various linkages of the different lysine residues. Proteins that are modified by the multiubiquitin chains linked via K48 or K29 are degraded by the proteasomes. Monoubiquitination or modification by K63 linked multiubiquitin chains do not promote proteasomal degradation, but rather serves to alter the function of a protein substrate or mediates novel protein-protein interactions (http://www.biochem.mpg.de/jentsch/Jentsch.html).

1.5 E3 Ubiquitin ligase

The E3 enzymes are required as an accessory factor by some ubiquitin conjugating enzymes for recognition of certain protein substrates. The E3 enzymes form complexes with specific ubiquitin conjugating enzymes. The E3 proteins are known to recognize the N-terminus of the substrate proteins, but there are other E3 proteins that are capable of recognizing internally located signals. Hence E3 enzymes can facilitate both substrate recognition function and may sometimes mediate ubiquitin conjugating reactions without the interaction of E2 enzymes (Jentsch, 1992).

1.6 The diversity of the E2 ubiquitin conjugating enzymes

The focus on E2 enzymes, is because they form the basis for further study in this project. First for the systematic organization of E2 enzymes into phylogenetic families, second for the structural and functional implications of family membership, third for the identification of active site signatures, as a means of understanding function and structure.

Thirteen different UBCs have been identified in yeast and at least 24 in human (see Table 1.1).

Table 1.1 Sub cellular localization and /or function(s) of UBC (E2s)

Yeast (Sc) & human UBC	Functions
ScUBC1	It is a ubiquitin conjugating enzyme that plays a vital role in the vesicle biogenesis and ER-associated protein degradation (ERAD) and is a component of cellular stress response. It also mediates selective degradation of short lived and abnormal proteins
HIP2	HIP2 binds selectively at the N-terminus of Huntingtin.
ScUBC2	It is a ubiquitin conjugating enzyme (E2) involved in the ubiquitin mediated N-end rule protein degradation (with Ubr1p), telomere silencing, sporulation, and in post-replication repair (with Rad18p)
UBE2A, UBE2B	The protein is strongly conserved in eukaryotic evolution, and plays an important role in various cellular processes.
ScUBC3	It is a ubiquitin conjugating enzyme (E2), which regulates cell cycle progression by targeting key substrates for degradation. It, together with Skp1p, Rbx1p, Cdc53p, and an F-box protein, forms a ubiquitin protein ligase called the SCF complex
UBE2R2, CDC34	Catalyzes the covalent attachment of ubiquitin to other proteins.
ScUBC4	It is a ubiquitin conjugating enzyme that is a component of cellular stress response, interacts with many SCF ubiquitin protein ligases. It mediates degradation of short lived and abnormal proteins.
ScUBC5	It is a ubiquitin conjugating enzyme that mediates degradation of short-lived and abnormal proteins and is a central component of the cellular stress response.

UBC enzymes	**Subcellular localization and /or function(s) continued:** **function(s)**
UBE2D1, UBE2D2,UBE2D3, UBE2E1,UBE2E2, UBE2E3	Catalyzes the covalent attachment of ubiquitin to other proteins and mediates the selective degradation of short-lived and abnormal proteins. It also functions in the E6/E6-AP-induced ubiquitination of p53.
ScUBC6	It is a ubiquitin conjugating enzyme that is involved in the ER associated degradation, which consists of the transmembrane segment at the end of the C-terminus, and is a substrate of the ubiquitin-proteasome pathway.
UBE2J1, UBE2J2	It functions in the selective degradation of misfolded membrane proteins from the endoplasmic reticulum, and catalyzes the covalent attachment of ubiquitin to other proteins.
ScUBC7	It is a ubiquitin conjugating enzyme that is involved in the ER-associated degradation pathway, which requires Cue1p for recruitment to the ER membrane.
UBE2G1, UBE2G2	Catalyzes the covalent attachment of ubiquitin to other proteins, may be involved in degradation of muscle-specific proteins.
ScUBC8	It is a ubiquitin conjugating enzyme that mediates the glucose-induced ubiquitination of Fructose-1,6-bisphosphatase (FBPase) and hence negatively regulates gluconeogenesis. It is a cytoplasmic enzyme that catalyses the ubiquitination of histones in vitro.
UBE2H	Catalyzes the covalent attachment of ubiquitin to other proteins, and is capable, in vitro, to ubiquitinate histone H2A.

UBC enzymes	Subcellular localization and /or function(s) continued: Functions
ScUBC9 UBE2I	It is a SUMO (small ubiquitin like modifier)-conjugating enzyme involved in the Smt3p conjugation pathway and also involved in the proteolysis mediated by the anaphase-promoting complexcyclosome (APCC). It is a nuclear protein required for S- and M-phase cyclin degradation and mitotic control. Catalyzes the covalent attachment of ubiquitin-like protein SUMO-1 to other proteins.
ScUBC10 UBE2S	It is a ubiquitin conjugating enzyme, which is essential for peroxisome biogenesis and it catalyses the covalent attachment of ubiquitin to other proteins. Catalyzes the covalent attachment of ubiquitin to other proteins.
ScUBC11 UBE2C	It is a ubiquitin conjugating enzyme, which catalyses the covalent attachment of ubiquitin to other proteins. Catalyzes the covalent attachment of ubiquitin to other proteins, and is required for the destruction of mitotic cyclins.
ScUBC12 UBE2M	It is a ubiquitin conjugating enzyme, which accepts the ubiquitin like protein NEDD8/RUB1 from the UBA3-ULA1 E1 complex and catalyses its covalent attachment to other proteins. The major substrate is CDC53/Cullin. Accepts the ubiquitin-like protein NEDD8 from the UBE1C-APPBP1 E1 complex and catalyzes its covalent attachment to other proteins, and is also involved in cell proliferation.

UBC enzymes	Subcellular localization and /or function(s) Functions
ScUBC13	It is a ubiquitin conjugating enzyme which is involved in the error free DNA postreplication repair pathway, DNA damage triggers distribution from the cytoplasm to the nucleus.
UBE2N	Plays a role in the DNA repair pathway and contributes to the survival of cells after DNA damage and has a role in the control of progress through the cell cycle and differentiation. It catalyzes the synthesis of non-canonical poly-ubiquitin chains that are linked through Lys-63, and mediates transcriptional activation of target genes. Ubiquitin-conjugating enzyme 13 (UBC13) has a key role in B-cell development and is important for both B-cell and macrophage activation (Yamamoto et al., 2006).
UBE2L1, UBE2L2, UBE2L3, UBE2L4, UBE2L6	Catalyzes the covalent attachment of ubiquitin to other proteins and mediates the selective degradation of short-lived and abnormal proteins. It also functions in the E6/E6-AP-induced ubiquitination of p53.
UBE2Q1, UBE2Q2	Catalyzes the covalent attachment of ubiquitin to other proteins

(Data from Table 1 was retrieved from
http://www.ensembl.org/Saccharomyces_cerevisiae/index.html;
http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl).

Table 1.2 Homologues of yeast UBCs that are UBC (E2) like proteins, and their functions illustrated

UBC likes	Functions
ScMMS2 CROC-1 (UBE2V1), UBE2V2	It has no ubiquitin ligase activity on its own. It catalyzes the synthesis of non-canonical poly-ubiquitin chains that are linked through Lys-63. This type of poly-ubiquitination activates IKK and does not seem to involve protein degradation by the proteasome. It plays a role in the activation of NF-kappa-B mediated by IL1B, TNF, TRAF6 and TRAF2. Mediates transcriptional activation of target genes. Plays a role in the control of progress through the cell cycle and differentiation, and DNA repair pathway and contributes to the survival of cells after DNA damage.
TSG101	TSG101 is a tumor susceptibility gene, which is important for growth restriction of normal cells. Mutations of the TSG101 gene are rare events in human breast cancers, but aberrant products from this gene are observed quite frequently. The occurrence of these abnormal products seem to be correlated with tumor and the mutation of another important tumor-suppressor gene called p53.

(Data from Table 1.2 was retrieved from http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl; Koonin and Abagyan, 1997)

Most of these enzymes in the Table 1.1 & 1.2 mentioned, appear to have a range of different cellular locations and/or substrate specificities. For example, yeast UBC6 and UBC7 bind to the endoplasmic reticulum and have been shown to be specifically involved in ERAD (Plemper and Wolf, 1999). In addition, both of these enzymes appear to have active sites, which are different from each other, and all other UBCs, with respect to their PROSITE signatures. For example, an entirely different PROSITE signature for yeast UBC6 homologues compared to all other UBC's has been described (see Figure 1.4). In multicellular organisms such as humans and plants, many of the homologues of the 13 different yeast UBCs appear to have been duplicated at least

once, e.g. there are 6 different homologues of yeast UBC4/5 in humans UBE2D1, UBE2D2, UBE2D3, UBE2E1, UBE2E2, UBE2E3 (see Table 1.1). In addition UBE2V1 and UBE2V2 appear to be homologue of yeast ScMMS2 (Table 1.1). So the UBC (E2s) represent a divergent population of homologous proteins.

1.7 The non–catalytic homologues of ubiquitin conjugating enzymes

The so-called noncatalytic UBCs, e.g. the tumour suppressor gene TSG101 and UBE2V [Li and Cohen, 1996; Li et al, 1997; Xiao et al, 1998] are an unusual class of proteins, which contain much of the conserved UBC domain but significantly lack the cysteine that is found in all UBC enzymes (Figure 1.4). The function of this class of proteins has been predicted to be a regulatory role, where the increased expression of this class of protein inhibits the ubiquitination of selected proteins by competing for protein substrates with other active enzyme, ubiquitin conjugating enzymes (Koonin and Abagyan 1997). However, this class of protein may still be able to interact with E3 ubiquitin ligases, and could therefore be actively involved in the ubiquitination process. In support of this latter hypothesis is the fact that, with the obvious notable exception of the highly conserved cysteine residue, this class of enzyme contains much of the UBC PROSITE signature, as well as other conserved ubiquitin domain amino acids at either side of the signature sequence. Therefore, activated ubiquitin may still be able to bind to this partial UBC active site before it is transferred to the active site of an E3 enzyme for catalytic addition to the target protein (Xiao et al., 1998).

1.7.1 TSG101

The tumor susceptibility gene (TSG101) contains a proline rich region and a predicted coiled coil domain, but the function could not be initially predicted from either the sequence or the structure. This suggests that TSG101 belongs to the group of apparently inactive homologues of ubiquitin conjugating enzymes (E2). When the protein database was searched at the NCBI to find the homologues of the TSG101 sequence using the BLAST2 program, a significant similarity to yeast and other homologues of E2 ubiquitin conjugating enzyme was found. Cysteine, which is within the active site of the E2s, was replaced by tyrosine in TSG101 (Koonin and Abagyan, 1997). TSG101 inhibits the ubiquitination and degradation of MDM2 and hence is directly involved in

cell cycle control by regulation of cellular p53 levels (Kloor et al., 2002). It could hence be concluded that TSG101and its UBC homologues have a common evolutionary ancestor, but they are most unlikely to possess the same function (Ponting, Cai and Bork, 1997).

1.7.2 CROC-1 (UBE2V1)

CROC-1 (UBE2V1) protein was isolated from human cDNA and is a homologue of yeast methyl methane sulphonate (MMS2) and the human MMS2 (UBE2V2). These homologues share significant amino acid sequence homology with that of CROC-1 and hence with the UBCs. The size and the predicted secondary structure of CROC-1 and MMS2 also resemble E2 enzymes, but they lack the active site cysteine residue which is highly conserved in the UBCs. It can be said that CROC-1 and MMS2 may form a separate UBC like protein family because though they resemble the UBCs, at the amino acid sequence level they are more homologous to each other than to the UBCs (Xiao, 1998). This is demonstrated by the multiple sequence alignment shown in Figure 1.4.

Figure 1.4 **Multiple sequence alignment of selected UBC6s with all other yeast UBCs and their selected homologues**

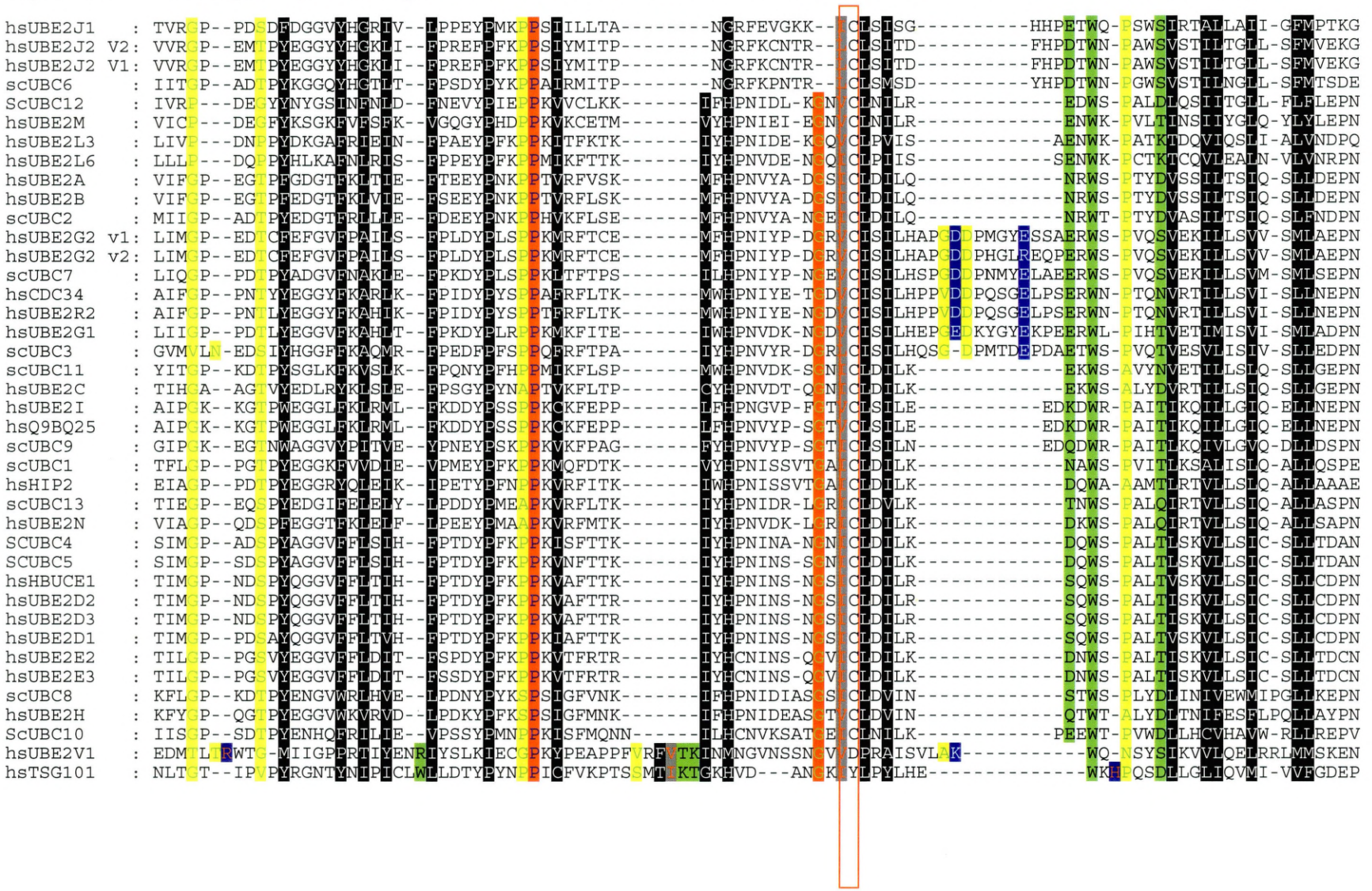


Figure 1.4 shows a CLUSTALX and GENEDOC generated multiple sequence alignment of yeast UBC6, it's two orthologous human UBE2J peptide sequences and all other yeast and human UBCs. The highly conserved cysteine, which is essential for this family of enzymes activity, is highlighted by a rectangle. This alignment also includes two human UBC related proteins, tumour suppressor gene 101(TSG101) and UBE2V1, which interestingly, lack the highly conserved cysteine active site residue, but as can be seen from the Figure, are still highly homologous to all other UBC's. From this alignment it can clearly be seen that with the notable exception of these latter named proteins together with the yeast UBC6 and its related human UBE2J sequences, all other known UBCs obey the following prosite signature [FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x-[LIV]. On analysing all known UBC6 related sequences which do not obey the typical UBC prosite signature, we have suggested the following prosite signature for UBC6: T-[PAR]-[NS]-G-R-F-x(3)-[KTE]-[RK]-[LIV]-C-[LMS]-[ST]-[IMF]-[ST]-x(2)-H-[PK]. It can clearly be seen from figure 1.4 that yeast UBC6 and the human UBE2J sequences obey this suggested signature. No PROSITE signature has yet been suggested for human TSG101 and its related proteins.*

1.8 UBCs AND UBIQUITIN LIKE PROTEINS (UBLs)

1.8.1 The relation of ubiquitin with SUMO and their functions

SUMO or "small ubiquitin like modifier" are distantly related proteins to ubiquitin with a 20% identity, which were first identified in mammals. SUMO is a Ubiquitin-like protein (Ubls) that conjugate to proteins, altering the properties of the modified protein and thus increasing the complexity of the proteome in eukaryotic cells. The mechanism of the SUMO conjugation is similar to that of ubiquitin conjugation involving the three enzymes E1, E2, and E3. The distinguishing feature of human UBC9 (UBE2I) that differentiates it from other conjugating enzymes for Ubls, is its ability to directly recognize substrate proteins. Hence the thioester UBC9-SUMO can catalyse formation of an isopeptide bond between the C-terminal carboxyl group of SUMO and the ε-amino group of lysine in the substrate protein (Hay, 2005). UBC9 has shown to be responsible for SUMO conjugation or sumoylation (Mo and Moschos, 2005). SUMO is required for DNA replication, repair and recombination and also to control the cell cycle progression (Hay, 2005; Desterro, Thomson and Hay, 1997).

1.8.2 Involvement of E1, E2, E3 in NEDD8

Another ubiquitin-like protein is NEDD8, which can be said to be the closest relative of ubiquitin. Although the enzymes involved in neddylation are the same as in the case of ubiquitin and ubiquitin-like proteins, the NEDD8 pathway involves only one E1, one E2 and a few E3s. Moreover NEDD8 involves only a small number of targets, in comparison to other ubiquitins that are involved in a greater number of targets. Human UBC12 (UBE2M) functions as a UBC for NEDD8, and has a 26 amino acid N-terminal extension upstream of its 150 amino acid residue conserved E2 core domain (Huang et al., 2004; Chiba, 2005).

1.9 UBCs and ERAD

The endoplasmic reticulum (ER) is the site in eukaryotic cells at which secretory proteins and membrane proteins enter the central vacuolar system which is made up of ER, Golgi apparatus, intermediate transport compartments, endosomes, lysosomes, and plasma membrane. The function of the endoplasmic reticulum (ER) is to properly fold in their native conformation and covalently modify the proteins which are inserted into the ER membrane or into the lumen before they arrive at their destination. Major ER modifications are N-glycosylation, disulphide bond formation and glycosyl phosphatidyl inositol (GPI) anchoring (Buschhorn et al., 2004). The folding process in the ER is controlled by a retention based quality control system consisting of ER-resident chaperones, protein disulphide isomerase (PDI), and lectins (Buschhorn et al., 2004). At the ER only properly folded proteins are allowed to pass through, whereas the misfolded proteins are eliminated. These misfolded proteins eventually are degraded in the ER lumen itself, and is known as ERAD or Endoplasmic Reticulum Associated Degradation (Plemper and Wolf, 1999). ERAD plays a central role during active secretion, cell growth, and normal turnover in eukaryotic cells (Mancini, Aebi and Helenius, 2003).

The three events that ERAD comprises are 1) substrate selection, b) transport to the cytoplasm, c) proteasome mediated proteolysis. It had been assumed before, that secretory proteins were degraded by ER resident proteases. It is now known that some ER substrates are re-translocated back to the cytoplasm and degraded by the proteasome although they are selected within the ER. The folding intermediates should be distinguished from the aberrant proteins, and for that the degree of substrate selectivity during ERAD must be high. The part of action of molecular chaperones is to assist in protein folding and is designed to recognize solvent exposed hydrophobic amino acid patches in unfolded proteins. The chaperones interact with ERAD substrates in yeast and mammalian cells, and for a number of ERAD substrates a correlation between substrate release from chaperones and proteasome mediated degradation has been observed. Chaperones help ERAD to fold a secretory protein, but if it does not fold, they retain the substrate in the soluble conformation. This prevents the formation of toxic aggregates and helps re-translocation through the ER membrane channel. If a protein folds and buries chaperone-associated hydrophobic patches, it will escape ERAD. If specific ER luminal chaperones are mutated in yeast,

ERAD substrates aggregate. Some misfolded subunits of multiprotein complexes, or misfolded proteins can remain bound to a chaperone indefinitely and are quite stable (McCracken and Brodsky, 2003).

There are various ERAD substrates that are associated with a disease for example the substrate AiPiZ (Z variant of the α1 proteinase inhibitor) is the most common genetic cause of liver disease in children and emphysema in adults. AiPiZ is a secretion-defective, accumulates in the ER, and is destroyed by ERAD. The most common substrate for the disease Cystic Fibrosis is ΔF508 CFTR. The mechanism of ERAD substrate recognition is not clear. It is thought to be likely that unfolded regions of aberrant polypeptides serve as recognition motifs that specify degradation, and that molecular chaperones play a role in targeting proteins to the ERAD pathway (Brodsky and McCracken, 1999). Selective proteolysis is an essential process in every cell (Biederer, Volkwein and Sommer, 1996).

It had also been said that when the mis-folded proteins are retained in the ER by chaperone association, an intra-lumenal timer is activated to select a protein for ER degradation. The degradation of some ERAD substrates requires the time dependent, enzymatic trimming of a terminal mannose residue on the middle chain of the N-linked core oligosaccharide. Proteins containing the shortened mannoside side chain (Mannose 8 instead of Mannose 9) are recognised poorly by glucosidase II, which trims a terminal glucose from the oligosaccharide. If the glucose residue is absent, calnexin binding does not occur, but if it is present, then glycoproteins bind calnexin, chaperones that retain mis-folded proteins in the ER and are known to play a vital role in ERAD. For mis-folded proteins, the decision to re-enter the calnexin cycle is made by the UDP-glucose: glycoprotein glucosyl transferase (UGGT), which adds glucose back to the oligosaccharide, which activates the calnexin re-association (McCracken and Brodsky, 2003). This is further illustrated in Figure 1.5.

Figure 1.5 Trimming of sugar chains and elongation of polyubiquitin that target the protein for degradation

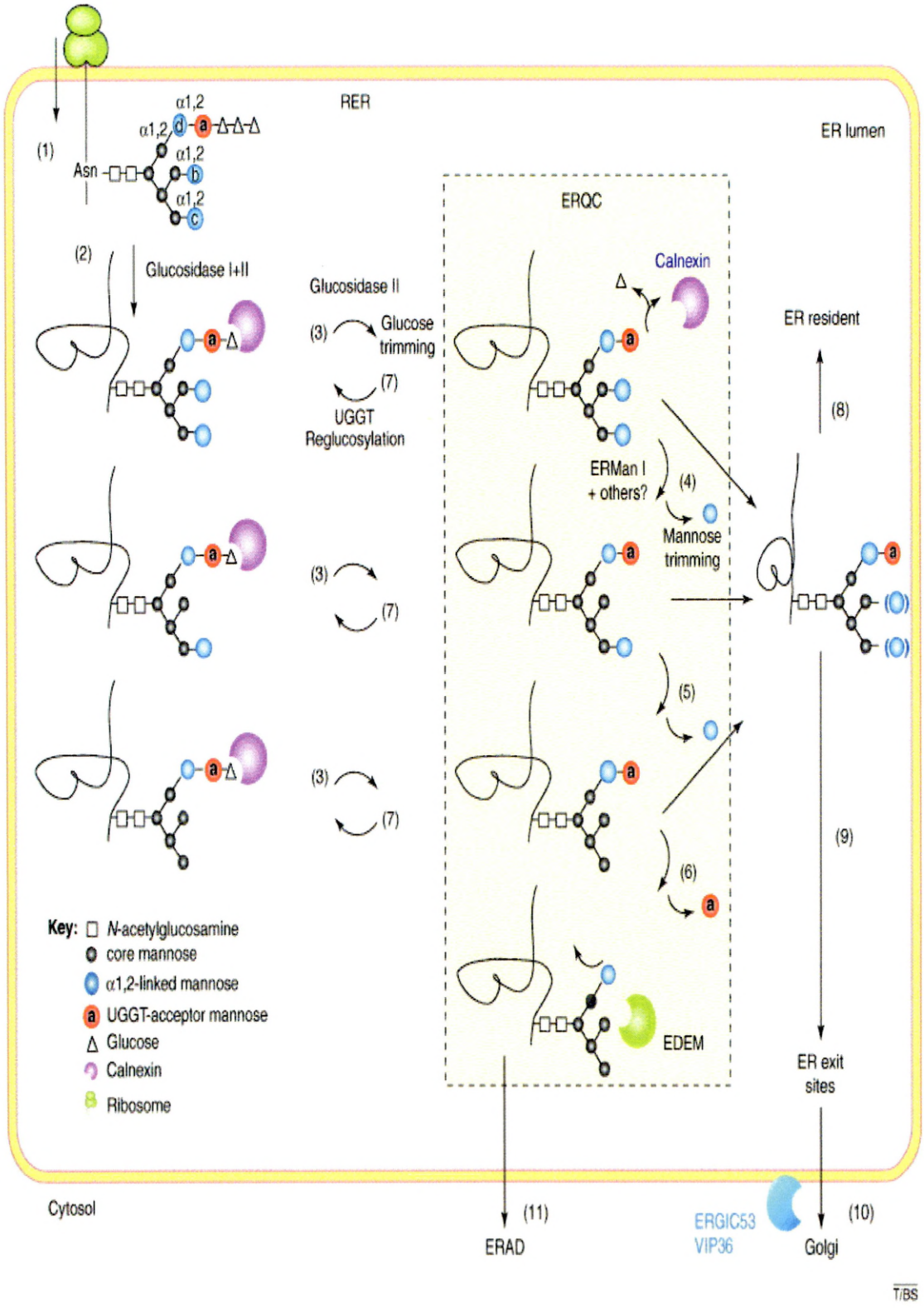


Figure 1.5 shows the step-wise trimming of mannose residues, which provides a timer for glycoprotein delivery to ERAD. (Step1) The oligosaccharide gets transferred to an asparagine residue in the mammalian rough ER (RER). (Step2) After that the two glucose residues gets trimmed which leads to binding of calnexin. (Step3-6) The calnexin folding cycles then starts, where trimming of the sugar residues takes place. (Step7) Reassociation with the calnexin takes place, followed by deglucosylation (step 3) and again mannose trimming. During any of these cycles the glycoprotein can acquire proper folding leading to its release from the cycle. (Step8) Once folded, the glycoprotein either remains as a stable ER resident or, (step9) is transported to the Golgi through the ER exit sites, where it is transported to its final destination, with the help of either ERGIC53, VIP36 or other lectins.

If even upon reaching step 6 the glycoprotein does not fold properly, the mannose-a residue is then removed, which leads to its release from the calnexin cycle, followed by the binding of the putative lectin EDEM, which leads to the ER degradation of the glycoprotein. The ERQC (ER quality control) compartment indicated, confirms some of the steps (Figure adapted from Lederkremer and Glickman, 2005).

The two ubiquitin conjugating enzymes (E2s) in yeast UBC6 and UBC7 are involved in the ERAD process. UBC6 is an ER membrane protein whereas UBC7 is not an integral membrane protein, but associates itself with an ER bound protein (see Table 1.1) called Cue1 (Tiwari and Weissman, 2001). Proteasomal degradation is not limited to proteins located in the nucleus and the cytosol; many transmembrane and lumenal proteins are degraded from the ER by the proteasomes, and this process of the proteasomal degradation of ER associated proteins is also known as ERAD. The steps involved in this form of ERAD are trimming of the N linked glycans, ubiquination, retrograde movement through the ER membrane, deglycosylation and degradation in the cytosol by proteasomes. A number of ERAD substrates are multiubiquitinated in yeast, which include the mutant forms of sec61, carboxypeptidase Y (CPY) and HMGCoA-reductase. In ERAD two yeast E2s, UBC6 and UBC7 have been implicated by genetic analysis, deletion of which stabilizes mutant sec61, CPY, Sss1p, Pdr5 and uracil permease. UBC7 protein associates with an ER bound protein, Cue1 as it lacks a membrane anchor, whereas UBC6 is a membrane protein, whose catalytic side faces the cytosol (Tiwari and Weissman, 2001).

There are two human homologues of yeast UBC6 namely UBE2J1 (NCUBE1) and UBE2J2 (NCUBE2). Similiarly there are also two human homologues of yeast UBC7 namely UBE2G1 and UBE2G2.

ER substrates can be translocated by another mechanism involving polyubiquitination. This involves the three ubiquitination enzymes E1 (ubiquitin activating enzyme), E2 (ubiquitin conjugating enzyme), and E3 (ubiquitin ligases) (McCracken and Brodsky, 2003; Taxis et al., 2003). UBC1 and UBC7 may also contribute to the polyubiquination of ERAD substrates, whose activities seem to represent the principal ubiquitin conjugation steps in this process (Friedlander et al., 2000). UBC6 has a central role in the ER Associated Degradation in yeast and other higher eukaryotes, where the misfolded protein is translocated out of the ER, ubiquitinated and degraded in the proteasome. This ubiquitination reaction is carried out by a complex, consisting of UBC6, UBC7 and Cue1 a membrane bound protein that couples the interaction of both UBC6 and UBC7 (Lester et al., 2000).

1.10 UBC6, PML-RARα and ERAD

PML-RARα is a fusion protein of promyelocytic leukaemia (PML) and the retinoic acid receptor-α (RAR-α), which causes acute promyelocytic leukemias (APL). The ubiquitination of N-CoR is stimulated by PML-RARα via UBC6 that is involved in the protein quality control. PML-RARα has two N-CoR interacting sites, which are required for the ERAD of N-CoR, which suggests that the abberant binding of PML-RARα to the N-CoR, may induce the ERAD of N-CoR (Khan et al., 2004; Atsumi et al., 2006).

1.11 UBCs, ERAD and human diseases

All mutated proteins are not degraded; only those mutations that cause a problem in protein folding are degraded. Misfolded proteins resulting from mutations are usually destined for protein degradation (Alfred et al., 2003).

Endoplasmic Reticulum Associated Degradation (ERAD) is the central element of the secretory pathway and also has the major implications for the generation of human diseases. There are two groups of disorders, the first being the result from the loss-of-function mutations in ERAD components that stabilize aberrant proteins which in turn accumulates and damages the cell, an example of it is the Parkinson's disease (see Table 1.3.2). The premature degradation of secretory or membrane proteins, is the second group of disorder, cystic fibrosis being an example (see Table 1.3.1) (Meusser et al., 2005; Plemper and Wolf, 1999). The following Tables 1.3.1 and 1.3.2 provide examples of diseases that are thought to result from defects in protein degradation.

Table 1.3.1 Diseases associated with defects in protein degradation

Cystic fibrosis
α_1 antitrypsin deficiency without liver disease
Congenital hypothyroidism
Thyro globulin deficiency
Thyroid peroxidase deficiency
Thyroxin binding globulin deficiency
Protein C deficiency
Disorders of lipid metabolism
LDL receptor defect
Lipoprotein lipase deficiency
Lipoprotein (a) deficiency
Hereditary hypoparathyroidism
Nephrogenic diabetes insipidus due to mutations in AVP receptor 2 or aquaporin-2
Growth hormone receptor deficiency
Osteogenesis imperfecta
Procollagen type I, II, IV deficiency
Albinism / Tyrosinase deficiency

Table 1.3.2 Diseases associated with defects in protein aggregation

Autosomal dominant neurohypophyseal diabetes insipidus
Liver disease in α_1 antitrypsin deficiency
Retinitis pigmentosa
Parkinson's disease
Alzheimer's disease
Huntington's disease

(Tables 1.3.1 & 1.3.2 were adapted from Rutishauser and Spiess, 2002; Aridor and Hannah, 2002).

ER associated degradation and cystic fibrosis

The Cystic Fibrosis gene encodes a large (around 168 KDa) epithelial membrane protein called Cystic Fibrosis Transmembrane Conductance Regulator (CFTR). Cystic Fibrosis is an autosomal recessive disorder affecting the lungs, pancreas, biliary tree, intestinal glands, reproductive organs and sweat glands (Gelman and Kopito, 2002). The CFTR functions as a regulator of chloride ion channel and also other ion channels.

Figure 1.6 Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) and its degradation pathway

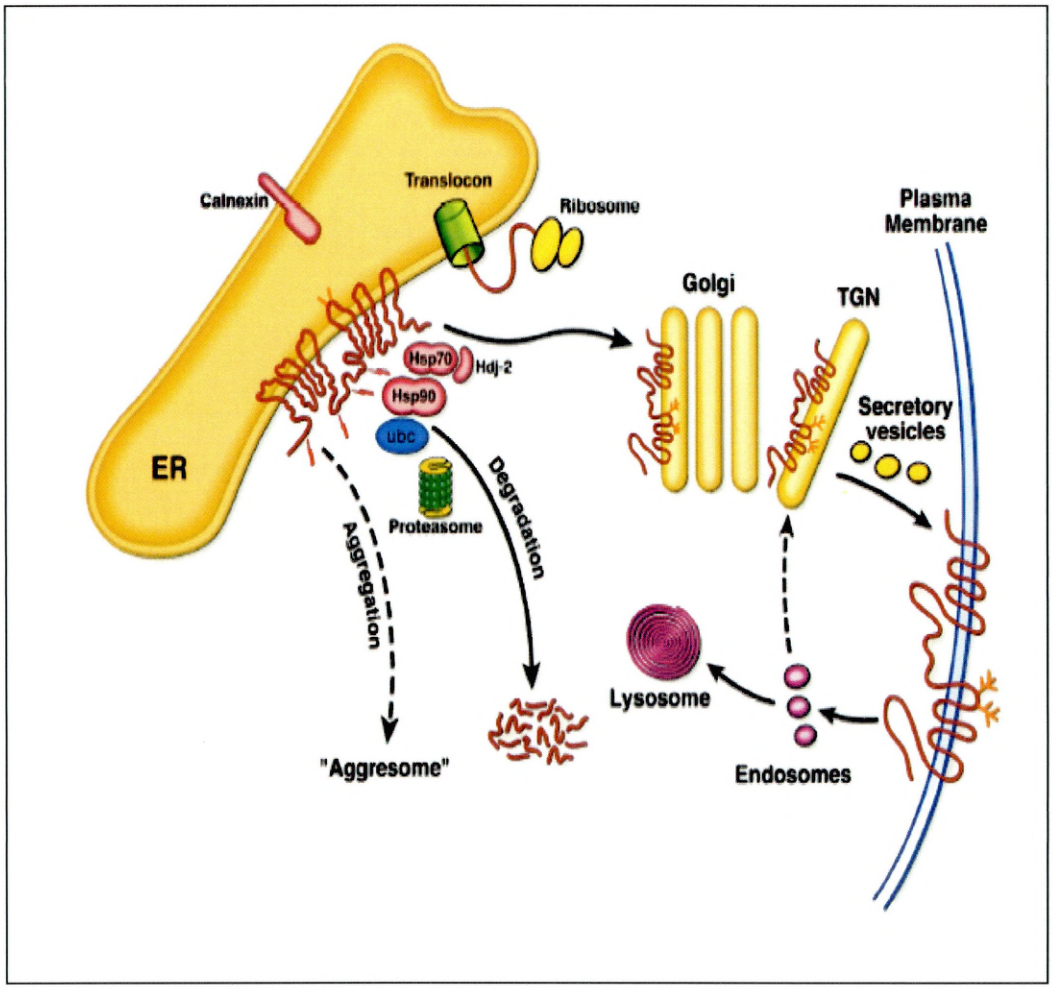


Figure 1.6 shows the fate of the CFTR molecule synthesised on the ER associated ribosomes. Calnexin binds to the attached core oligosaccharide chains, and in addition to it the Hsp90, Hsp70 and Hdj2 binds, where ubiquitination may occur. The completely folded CFTR is prevented from degradation, but those molecules that are misfolded are degraded by the proteasome. But when degradation is prevented, extensive aggregation of molecules occur that are not competent to be exported (Riordan, 1999).

The most common CFTR mutation found in most patients is a deletion of phenylalanine at position 508 (ΔF508)
(Rutishauser and Spiess, 2002; Ward, Omura and Kopito, 1995).
UBC6 and UBC7 are the E2s that function with the E3s such as gp78 and Doa10 on the cytoplasmic face of the ER to ubiquitinate a number of substrates. Therefore it is possible that E2-E3 complexes that contain UBC6 and/or UBC7 cause the degradation of CFTR ΔF508 (Younger et al., 2004). Human homologues of yeast UBC6 (UBE2J1 and UBE2J2) have recently, both been shown to be involved in the ubiquitination and proteasomal degradation of the most common misfolded mutant ΔF508CFTR. It had been found that overexpression of UBC6 and not UBC7, modulates the rate of CFTR ΔF508 degradation (Lenk et al., 2002). When the effects of the dominant negative forms of UBC6, UBC7, UbcH5A (UBE2D1), had on CFTR and CFTR ΔF508 expression were compared, it was found that UBC6, Ubc5A (UBE2D1) drove the accumulation of the B form of CFTR and CFTR ΔF508, whereas UBC7 had no apparent effect. This suggests that UBC6 plays a clear role in CFTR quality control. The specific inhibition of the UBE2J1 and UBE2J2 enzymes may therefore have therapeutic potential in the treatment of cystic fibrosis, as this would in turn inhibit the ubiquitination and subsequent degradation of ΔF508CFTR protein by the proteasome (Younger et al., 2004; Brodsky and McCracken, 1999; Yoshida et al., 2002).

1.12 UBC's as potential drug targets

As the ubiquitin system plays a central role in basic cellular processes, development of drugs that modulate the activity of the ubiquitination system, could provide a target oriented, highly specific approach to many diseases. The clinical importance of inhibiting the enzymes controlling the ubiquitination pathway (which may affect many processes non-specifically), may depend on striking an acceptable balance between the beneficial effects, the toxicity and the deviation of the treatment. The beneficial effects in certain diseases like Cancer, Asthma, Brain Infarct, Cystic Fibrosis, Parkinson's, has been strongly suggested from experimental evidence. Overexpression of UBC6 but not UBC7, modulates the rate of CFTRΔF508 degradation. Here UBC6 could act as a potential drug target in altering the degradation function of the CFTRΔF508. Specific regulators of this proteolytic pathway should thus be valuable therapeutic tools for treatment of several severe human diseases (Bence, Sampat and Kopito, 2001; Mira et

al., 2004; Kostova and Wolf, 2003; Plemper and Wolf, 1999; Ciechanover and Schwartz, 2002; Kaiser and Huang, 2005).

The nuclear receptor corepressor (NCoR) and the related factor known as silencing mediator of retinoic acid and thyroid hormone receptor (SMRT) are essential components of multiprotein complexes, which act to repress transcription of nuclear hormone receptors, such as the thyroid hormone receptor (TR) and retinoic acid receptor, in the absence of their appropriate ligand (Ogawa et al., 2004). They play a crucial role in transcriptional repression by multiple classes of transcription factors, including some nuclear hormone receptors (Frasor et al., 2005). PML-RARα stimulates the ubiquitination of N-CoR via UBC6, which is involved in the protein quality control. As UBC6 is involved in the ER degradation and causing the disease acute promyelocytic leukaemia (APL), it is again, one of the potential drug target in the control of the disease (Khan et al., 2004).

N-CoR has been seen to be regulated by ubiquitin-mediated protein degradation by an interaction of E3 ubiquitin ligases Siah2, which leads to the ubiquitination and degradation of N-CoR by the 26S proteasome. Estrogen down regulates the N-CoR protein levels in estrogen receptor (ER)-positive breast cancer cells without affecting N-CoR mRNA levels and the corepressor SMRT levels (Frasor et al., 2005). Here the interaction of E3 ubiquitin ligase Siah2 has been mentioned, but there could be a possibility of the interaction of an E2 (Ubiquitin conjugating enzyme), which by affecting the N-CoR level could cause breast cancer. Hence, here also UBC6 could act as a potential drug target in the control of breast cancer.

1.13 Phylogenetic and structural studies to determine the function of (E2s) UBC orthologues

As can be seen in the previous sections, UBC (E2s) are involved in diverse cellular processes and have been shown to possess different functions. For example yeast UBC6 and 7 have been shown to be specifically involved in misfolded proteins that are targeted for ERAD (Plemper and Wolf, 1999; Kostova and Wolf, 2003) while human UBC9 has been shown to be involved in the specific ubiquitination of P53 helping to control its vital cellular function. Interestingly the active sites of yeast UBC6, yeast UBC7 and human UBC9 are significantly different. For example UBC6 has amino acids G, R, F, which is absent in all other UBCs, whereas UBC7 has 13 extra amino acids

which are not found in any other UBCs except UBC3. Despite its possible therapeutic importance as mentioned in the previous section, the structure of yeast UBC6 and its human orthologue (UBE2J1) remain to be determined.

1.14 Structure of UBCs (E2's)

As the ubiquitin pathway involves the interaction of ubiquitin conjugating enzymes with different substrates, or other accessory proteins, distinct specificity to individual enzymes could be conferred by specific structural features. Hence it is essential to know the three dimensional structure of these ubiquitin conjugating enzymes (UBCs). There are 13 different yeast UBCs, each having different functions and for which some of the UBC structures have been resolved. Examples are the structure of *Arabidopsis* UBC1, *Saccharomyces cerevisiae* UBC4, *Saccharomyces cerevisiae* UBC7 and human UBC9 (Cook et al., 1997). As UBC7 is thought to ubiquitinate and hence cause the degradation of the common mutant form of the CFTR protein (i.e. ΔF508), the structural interpretation of UBC7 is essential for a proper understanding of the disease. Before the elucidation of the crystal structure of UBC7, two other ubiquitin conjugating enzyme structures (AtUBC1 and ScUBC4) had been determined (Cook et al., 1992; 1993). UBCs have been broadly divided into four classes of enzymes, based on amino acid sequence comparison. Class I enzymes consists of a relatively conserved catalytic core domain of around 150 amino acid residues with having at least 25 % sequence identity. The class II and III enzymes have an external attachment to the core domain, of either a C-terminal or N-terminal extension respectively. The class IV enzymes have both the C-terminal and N-terminal extension. These extensions to the core domain confer specificity to a certain degree for enzyme-substrate recognition, or provide a localization signal. Residing within the core domain itself are both the specificity and localization signals. Either the single member substrate or the multiple member substrates from the same or different classes, are involved in different cellular processes, and hence comprise distinct functional subfamilies. (Tong et al., 1997; Jentsch, 1992).

The determined structures of UBC1 & UBC4 are both class I enzymes and consists entirely of the conserved core domain. By comparing these two structures it could be said that class I enzymes are highly conserved in their three-dimensional folding state. As most of the identical residues of the two enzymes are either buried or clustered on the surface that lies next to the ubiquitin accepting cysteine, variations are found (Cook

et al., 1997). It has been speculated before that the function of the conserved surfaces is to help in protein-protein binding during ubiquitin thiol ester formation and may be the E1 binding site of all E2 enzymes. The surfaces which are not conserved may enable an individual Class I enzyme, to interact with its respective substrate or E3 protein.

Further structural studies have been carried out of UBCs in order to find out whether any UBC contain more than the minimum core domain. The structure of UBC7, a Yeast UBC (E2), was determined, which contain two insertions relative to the other E2s (UBC1 and UBC4) (Cook et al., 1992; 1993; 1997). Yeast UBC7 has a molecular weight of 18521 Dalton and has 165 amino acids. Deletion of the UBC gene causes a cadmium hypersensitivity phenotype (Jungmann et al., 1993). UBC7 protein has functional implications that have been confirmed at the genetic level, as it is involved in the ubiquitin dependent degradation of a large group of proteins. The three dimensional structure of the yeast UBC7 protein has an overall folding, which is similar to the UBC1 and UBC4, although it has 38% sequence identity to *Arabidopsis* UBC1 and 37% with that of yeast UBC4. These three structures have been compared which revealed a surface region that could be the E1 binding site which is common to all E2 enzyme. There are surface regions, also which are divergent, that could enable an individual class I enzyme to interact with its respective substrates or E3 (Cook et al., 1997).

Figure 1.7 Multiple sequence alignment of UBC7, UBC4 and UBC1

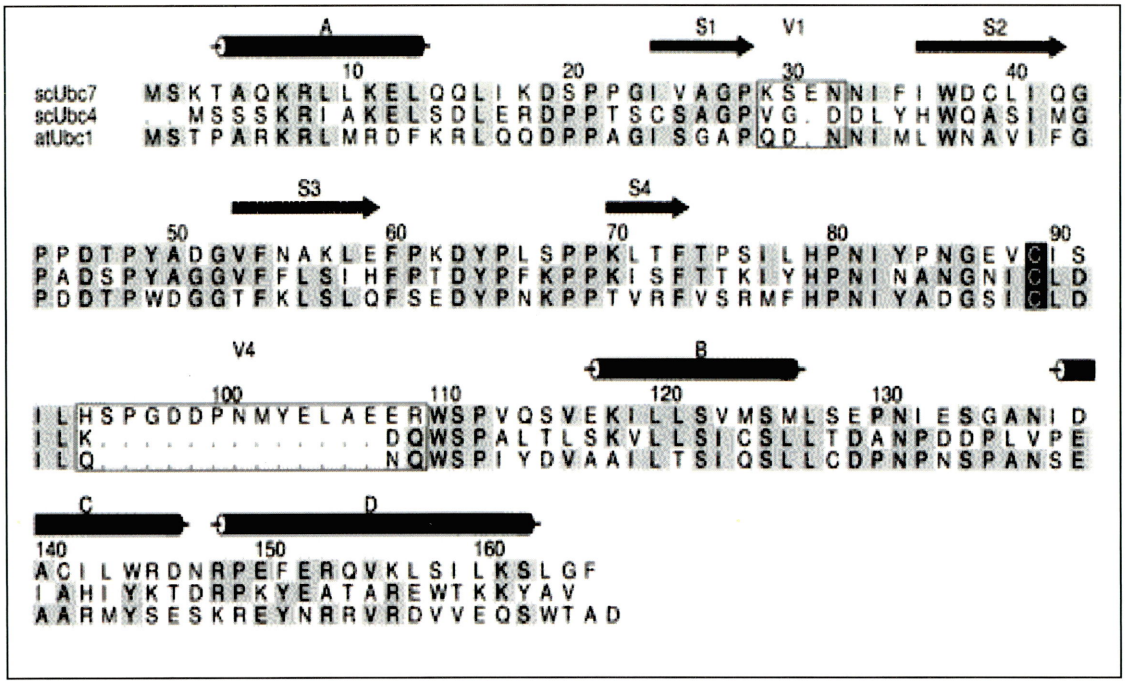


Figure 1.7 shows the multiple sequence alignment of UBC7, UBC4, and UBC1. The position numbers and the secondary structural features that are shown are of UBC7. The active site residue cysteine residue is shown with a black background. From the multiple sequence alignment it can be seen that UBC7 has 14 extra amino acids, which are not found in either of the UBCs shown by the regions indicated V1 & V4. Regions S1 to S4 are the four antiparallel β sheets and regions marked A, B, C and D are the four helices (Cook et al., 1997).

The 14 extra amino acids are in two separate regions V1 and V4 of Figure 1.7. One of the extra amino acid in region V1 is Glu 31 and the other 13 extra amino acids are Ser 95 – Glu 107 which is just next to the ubiquitin-accepting cysteine at position 89. These two predicted insertions though have altered the two surface regions, have not altered the overall folding of the pattern of Class I enzymes as in UBC4 and UBC1. The overall dimension of the asymmetric UBC7 molecule is approximately 24 Å x 41 Å x 49 Å. The figure below shows the α-carbon backbone of the UBC7 molecule.

Figure 1.8 Stereo diagram of UBC7 on the α-carbon position

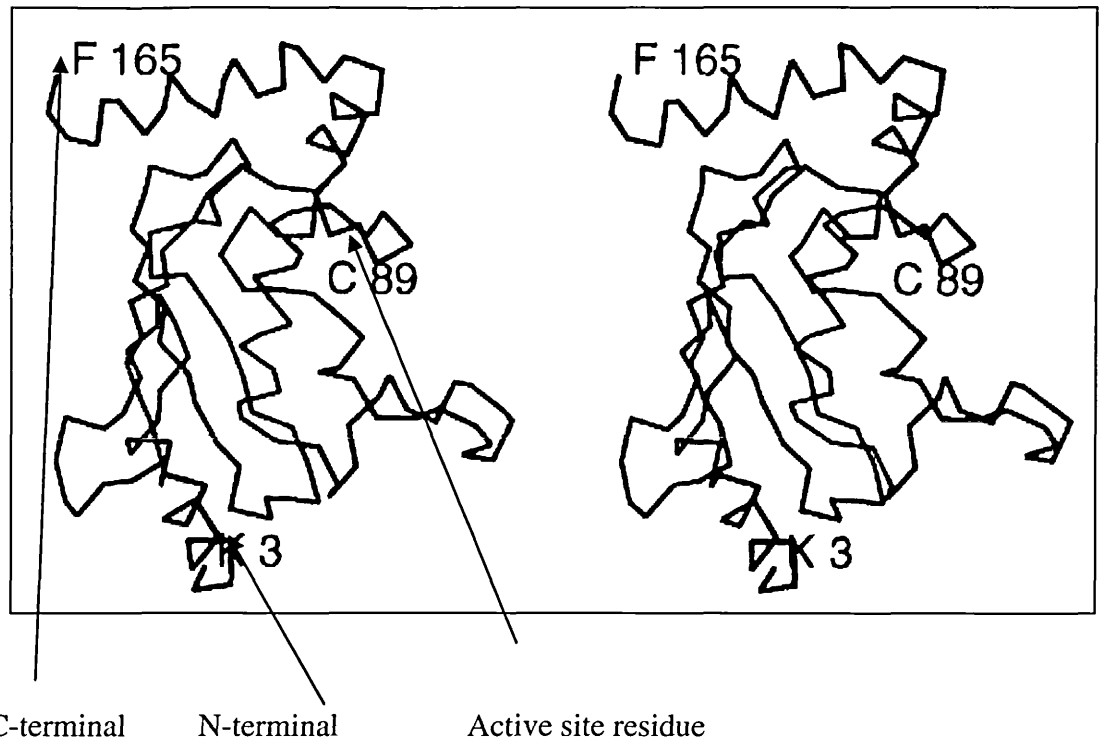
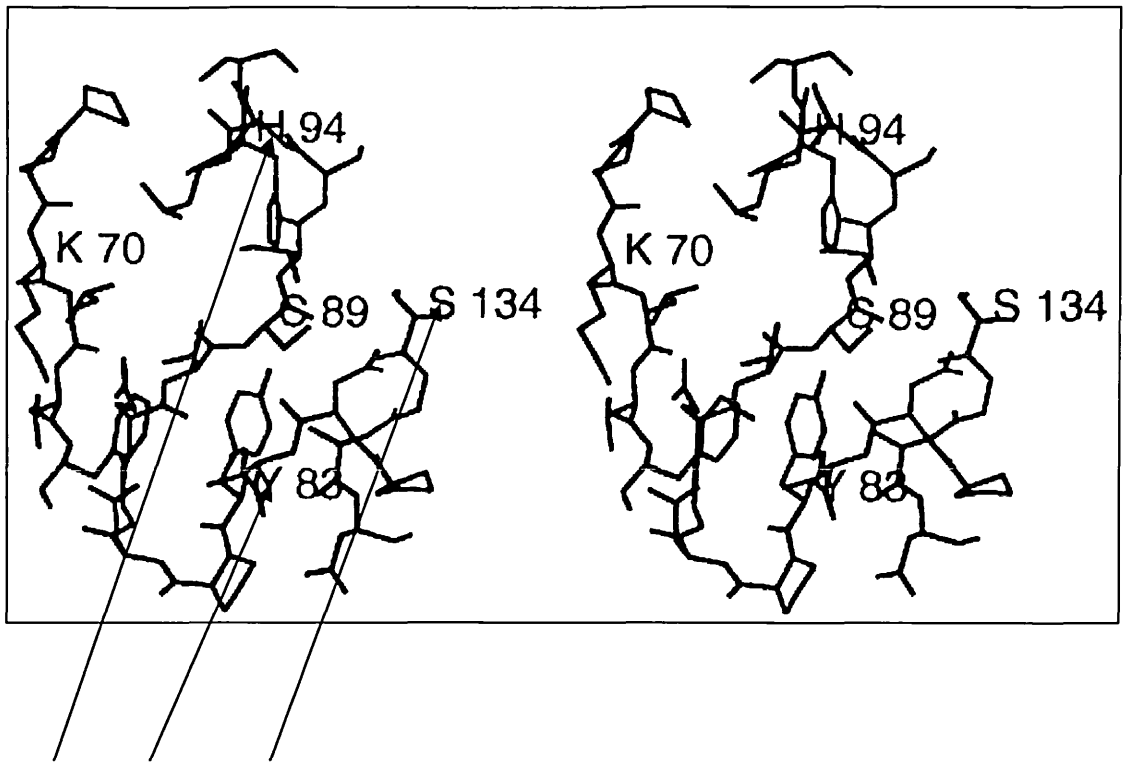


Figure 1.8 shows the stereo diagram on the α-carbon positions. The ubiquitin accepting cysteine (C89) and the N-(K3) and C-(F165) terminals are labelled (Cook et al., 1997).

UBC7 molecule has one antiparallel β-sheet like that of UBC1 and UBC4, with four strands (labelled S1 to S4 of the Figure 1.7), bounded by four helices on each end and on one side (labelled A, B, C, D). On the surface of the protein there is one face of the β-sheet and the residues of the four β-sheets include from 24-27, 36-42, 53-59, 71-73. Residues 7-13, 116-128, 138-146, 148-162 are the four α-helices and there are two hydrogen bonds to strand S4: Phe73 N- Gly86 O and Leu 71 O- Val88 N. What is not present in either UBC4, nor in AtUBC1 are the three short stretches of 3_{10} helix in UBC7 (which are residues 91-93, 104-108, and 132-134). Located in the long extended stretch between the fourth strand (S4) of the β-sheet and the second α-helix (B) of the UBC7 molecule is Ubiquitin-accepting cysteine.

Figure 1.9 **Stereo diagram of yeast UBC7**



Residues at the loop regions surrounding the active site residue cysteine (C89).

Figure 1.9 is a stereo diagram of the environment around the ubiquitin-accepting cysteine in yeast UBC7. The residues 68-73, 80-95, and 134-137 are shown (Cook et al., 1997).

Figure 1.9 shows the stereoview of the environment around the ubiquitin-accepting cysteine. Similar to ScUBC4 and AtUBC1, ScUBC7 also has a large insertion close to Cys89, the cysteine side chain is exposed and sits in a slight depression on the surface. Surrounding Cys89 are three loops, the closest residues are Tyr83, His94, and Ser134. Closer to the Ubiquitin-accepting cysteine in UBC7 is the loop containing Ser134, compared to the corresponding loop in yeast UBC4.

The ribbon diagram of the superimposed yeast UBC4 and UBC7, showing the conserved tertiary folding of class I enzymes is as follows:

Figure 1.10 Superimposed structures of UBC7 and UBC4

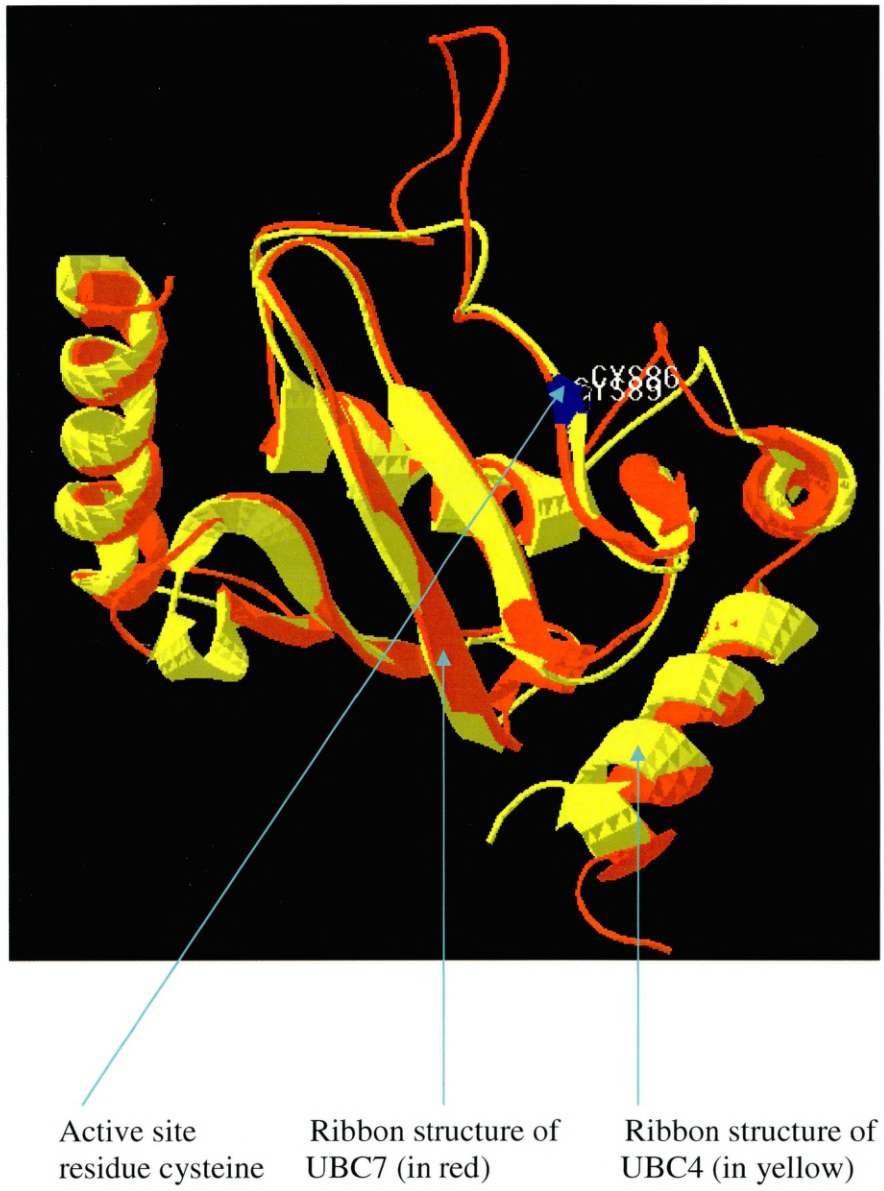


Figure 1.10 is the ribbon structure of yeast UBC7 molecule (in red), superimposed on the yeast UBC4 molecule (yellow). The ubiquitin accepting cysteine is marked in blue and various other residues like 3-31, 33-94, 108-164 from UBC7 and residues 1-29, 30-91, and 92-148 from UBC4 are superimposed (Cook et al., 1997). This Figure has been reproduced in the software called DeepView.

The remaining α-carbon atoms of UBC7 excluding residues in the two insertion, can be superimposed with the corresponding atoms in UBC4 with a root mean square deviation of 1.8 Å. The superposition of ScUBC7 with AtUBC1 gave a similar root mean square value of 1.8 Å. As the superposition of ScUBC4 and AtUBC1 gives an RMS value of 1.6 Å (Cook et al., 1993), there is not much of difference in the RMS values. The differences occur mostly at the loops but they are quite small. All the four strands of the β-sheets in ScUBC7 are slightly shorter as compared to that of ScUBC4. Compared to the 27 residues in ScUBC4 there are 21 residues in the four β-strands of ScUBC7. From the deviations in the superposition, it can be judged that the helix A and the loop linking it to the first β-strand in ScUBC7 are very much similar to the corresponding regions in AtUBC1 than in ScUBC4. Superposition at the C-terminal end of helix D compared to both AtUBC1 and ScUBC4 shows significant deviations. Cis-Pro62 of ScUBC4 and cis-Pro64 of AtUBC1 aligns with the cis-Pro65 of ScUBC7 as shown in the previous alignment Figure 1.7

The most significant difference between UBC7 and all other UBC E2s are the regions of extra amino acids as indicated as V1 and V4 of the alignment Figure 1.7. They represent hypervariable surfaces in a common tertiary fold, when these three crystal structures are compared in these regions. By the insertion of the extra 13 amino acids in the UBC7, the effect is that a large surface loop has been introduced flanking to one side of the Ubiquitin-accepting cysteine. This loop in UBC4 is comprised of Lys91-Gln93, whereas this region in UBC7 is comprised of 16 amino acid residues (His94-Arg109), and contains two tight turns and a short stretch of 3_{10} helices.

Figure 1.11 shows the multiple sequence alignment of amino acids of the 13 yeast E2s around the active site cysteine.

Figure 1.11 Alignment of all 13 yeast UBCs around the active site residue region

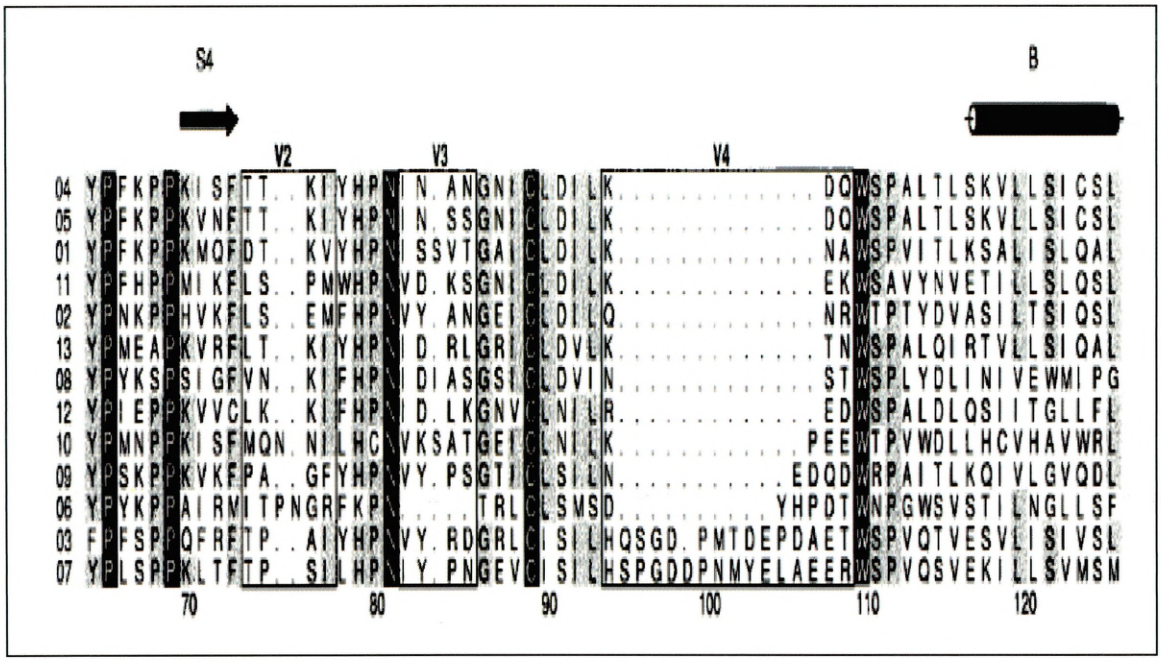


Figure 1.11 is the alignment of all 13 yeast UBC sequences around the active site cysteine. The invariant residues are highlighted in black background and the identical ones are in gray. The secondary structural features above the alignment and the numbers below refer to that of UBC7 sequence. Regions marked V2, V3 and V4 are the variable regions of UBC7, and region marked S4 is the fourth strand and region marked B is the helix of the UBC7 peptide sequence (Cook et al., 1997).

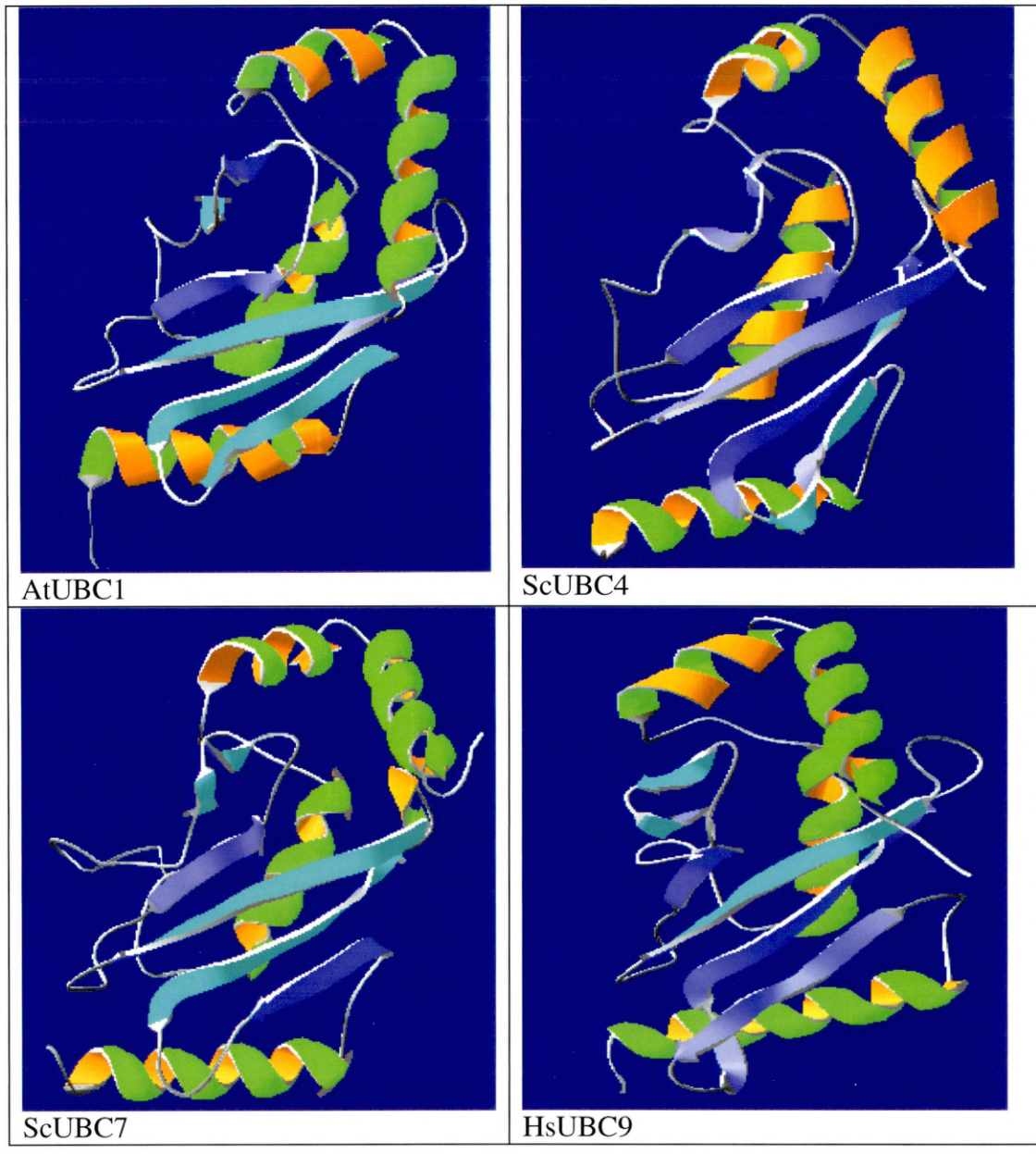
Eight of the thirteen E2 sequences are similar in the loop region V4 by consisting of 3 residues as can be seen in Figure 1.11. The remaining five sequences consist of a varying length of 4-16 amino acids in this loop region V4. It is evident that even though UBC7 has the 13 extra amino acids, the regions adjoining this loop in UBC7 are very much similar to that of UBC4. This loop in UBC7 is flanked on the C-terminal side by a tryptophan (Trp110) and by a highly conserved leucine (Leu93) on the N-terminal side. When the UBC4 and UBC7 has been superimposed it is seen that the α-carbon atoms at the position of Leu93 and Try110 deviate between the two molecules with values of 0.14 and 0.30 Å respectively, which is much below the root mean square deviation (RMSD) value of 1.8 Å. Yeast UBC7 has a homologue in C. elegans, which has a similar 13 residue insert. A similar 12 residue insert is present in yeast UBC3 which also catalyses the formation of Lys48-specific ubiquitin-ubiquitin ligases. The ability of

UBC3 to catalyse the formation of specific ubiquitin-ubiquitin linkages at Lys48 is not possible by any other UBCs (Banerjee et al., 1993). In the previous multiple sequence alignment of Figure 1.11, the region of V1 of an extra amino acid, in the turn between the first two β-strands of UBC7 and UBC4 has the effect of shifting the tight turn by two residues, which lies at the surface of the molecule and protrudes slightly. There are at least two other regions in yeast E2 sequences that may contain variable length insertions, which are located between the last β-strand and the ubiquitin accepting cysteine (V2 and V3 of the alignment Figure 1.11. These insertions are absent in both the UBC4 and UBC7, which exhibit deviation in the superimposed structures. The four variable regions that are highlighted in the yeast UBC7 are not randomly distributed, but are located on one surface, which is quite distinct, as they are randomly distributed in other E2s.

From the detailed structural discussion of the yeast UBC7, its amino acid sequence comparison with yeast UBC4 and plant UBC1, and also its structural comparison with yeast UBC4, it is found that the structure of yeast UBC7 is quite distinct to that of all other UBCs, especially with the 13 amino acid insert in UBC7.

The structural basis by which ubiquitin conjugating enzymes (E2s) determine substrate specificity remains unclear. It was found that the extra 13 amino acids loop can play a role in substrate specificity in UBC7. When the sequences of many E2s were compared to this extra amino acid region, it suggests that it may represent a hypervariable surface in a common E2 tertiary fold (Lin and Wing, 1999). Thus a functional characterization was given to the yeast UBC7 from its determined structure. Figure 1.12 illustrates selected UBC (E2s) whose three dimensional structures had already been determined.

Figure 1.12 Some UBC (E2s) whose structures have already been determined



Listed in Figure 1.12 are some of the UBCs (Arabidopsis thaliana UBC1, Saccharomyces cerevisiae UBC4, Saccharomyces cerevisiae UBC7 and human UBC9, whose structures had already been determined by X-ray crystallography. All of these UBCs have their active site residue as cysteine and consist of four Alpha helices and a four stranded antiparallel beta sheets (Cook et al., 1992; 1993; 1997; Tong et al., 1997).

The structure of all the UBCs mentioned in the table 1.12, i.e. AtUBC1, ScUBC4, ScUBC7 and HsUBC9, are derived from the coordinates of their X-ray diffracted crystal structures.

The commonality in all these structure is that all belong to the Class I group of enzymes where the catalytic core domain of about 150 amino acids is conserved. All of these mentioned UBCs consist of four Alpha helices and a four stranded antiparallel Beta sheet. All have their active site residue cysteine. The overall similarity of the human UBC9 structure with that of the plant UBC1 and yeast UBC4 suggests that the folding of the catalytic domain of the family of UBC enzymes is conserved in all eukaryotes.

Despite such similarities in the above mentioned structures, there are considerable differences in the catalytic active site region. There are 10 residues in UBC9 within 6Å of the sulfhydryl group of the ubiquitin accepting cysteine. Cys93, Asn85, Tyr87, Lys101, and Asp127 are among those 10 residues that most likely mediate the catalytic activity. Only Asn85, Leu94, and Pro128 are conserved compared with UBC1 and UBC4. Most of the difference between the two structures of UBC1 and UBC4 occurs in the loop region. Unlike UBC1, only the first two residues at the N-terminus are non helical, and there is no tight turn at the beginning of the helix. The major structural difference between these two structures, lie in the loop region of (14-21 and 41-46 of UBC4) and residues (16-23 and 43-48 of UBC1). Moreover there is greater difference at the amino acid sequence level of UBC6 to that of all other UBCs. The catalytic active site region has amino acids that are unique to UBC6, like G, R, F. To study in details the three dimensional structure difference of UBC6 to that of all other UBCs, it is necessary to derive the structure of UBC6 (human UBE2J1 and UBE2J2).

The structures of yeast UBC7 and human UBC9 have been determined and compared to other UBC structures. Despite its possible therapeutic importance the structure of yeast UBC6 and its human orthologue (UBE2J1) remain to be determined.

The X-ray crystallographic analysis procedure adopted in determining the structures of all the UBCs mentioned above has given a good impetus to carry out the crystal structure determination of UBE2J1.

CHAPTER TWO

AIMS AND OBJECTIVES

2. Research aims and objectives

This project firstly aimed to construct a functional phylogenetic tree for UBCs in selected eukaryotic species, whose genomes had been fully sequenced. All yeast UBCs whose functions and in some cases structures have previously been experimentally determined were included in this analysis in order to try to construct a phylogenetic tree, with different branches representing the different orthologous functions of UBCs. This phylogenetic tree would also importantly, help in sorting out the nomenclature naming of all UBCs, especially the human UBCs.

Secondly this project aimed to try to determine the 3 dimensional structure of the active site region of human UBE2J1, using both X-ray crystallography and by computational methods. It is hoped by the resolution of this structure, specific inhibitory drugs may be designed to stop the degradation of mutant CFTR protein and alleviate the symptoms of cystic fibrosis in affected patients.

Lastly this project aimed to determine the previously unknown UBC active site for the *Drosophila* transcription factor $TAF_{II}250$.

CHAPTERS 3 – 6
INTRODUCTION TO METHODOLOGIES USED IN THE PROJECT

CHAPTER THREE

INTRODUCTION TO PHYLOGENETIC ANALYSIS

In order to carry out Phylogenetic analysis to try to statistically determine protein or DNA sequence evolutionary relationships, one has to first try to construct a meaningful multiple sequence alignment (MSA).

3.1 Multiple sequence alignment (MSA)

3.1.1 Introduction of MSA

Multiple sequence alignment is the key component of biological sequence analysis techniques. There are few procedures in bioinformatics that do not require at one point or the other the multiple sequence analysis. The multiple sequence analysis helps in 1) the identification of the PROSITE signature pattern of a protein, 2) the building of a domain profile needed for identifying the most remote member of a protein family, 3) structure prediction, 4) and phylogenetic analysis. Unfortunately it cannot always be statistically determined which particular multiple sequence alignment is the best. It is often preferable when carrying out an MSA to try analysing in different MSA algorithms to see which alignment looks most biologically correct. Following the automated MSA analysis, further manual alignments may be necessary to align small biologically significant motifs (Poirot, O'Toole and Notredame, 2003; Briffeuil et al., 1998)

3.1.2 MSA tools

In multiple sequence alignment, a sequence is aligned with another sequence, which is most similar to it, and then successively aligns with the next most similar ones. There are a number of multiple sequence alignment programmes for example T-Coffee (Notredame, Higgins and Heringa, 2000), CLUSTALW (Higgins et al., 1994), CLUSTALX which is a graphical interface of CLUSTALW, PILEUP (Dolz, 1994), EMBOSS (Rice, Longden and Bleasby, 2000). Generally two basic classes of alignment programs are used, one is global alignment programme, which aligns the sequence along their whole length, and the other is local alignment programme, which aligns sequences only around the conserved regions. Among the two, the most accurate and reliable alignment of equidistant sequences, sequences of divergent families, and orphan sequences with a family, is the Global alignment. CLUSTALW uses the global

alignment program to construct an alignment along the whole length of sequence (Thompson, Plewniak and Poch, 1999). Following the successful construction of an amino acid MSA (e.g. see Figure 3.2) shading of the same amino acids, or amino acids of similar property can help identify structurally similar protein domains. Figure 3.1 shows the properties of the amino acid residues.

Figure 3.1 Venn diagram of the properties of the amino acid residues

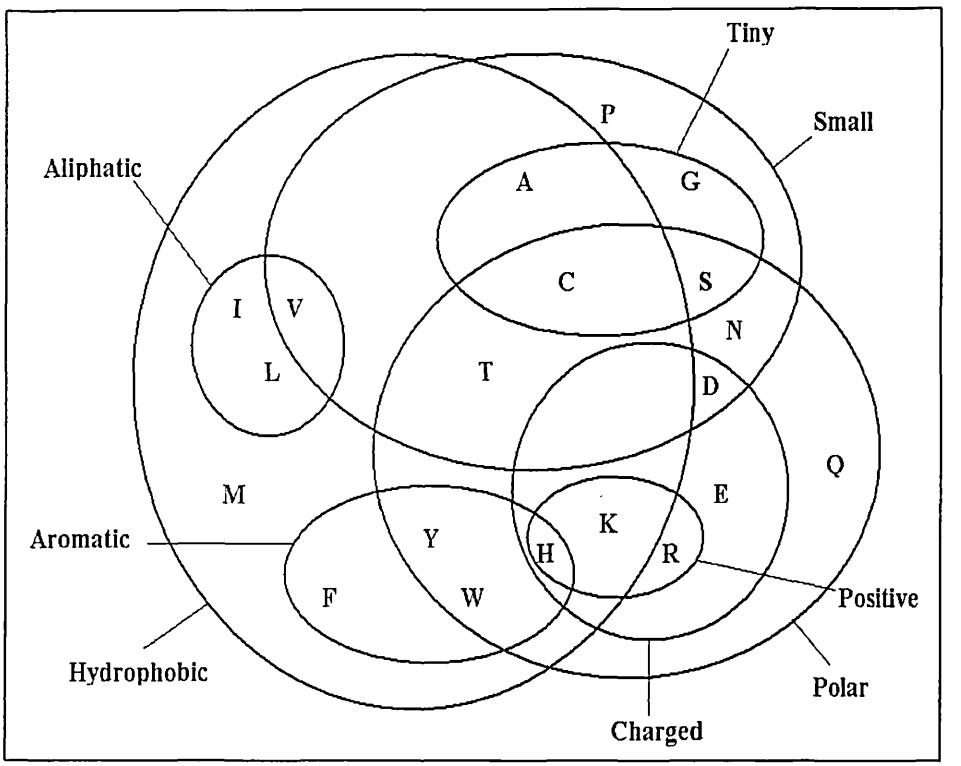


Figure 3.1 adapted from (Sternberg, 1996). The Venn diagram above shows the relationships of the 20 naturally occurring amino acids, on the basis of their physio-chemical properties. There are two major sets: one having amino acids that have a polar group, and the other containing hydrophobic amino acids. There are nine amino acids in the set "small", and within it there is a smaller set called "tiny" having amino acids which have at the most two side chain atoms. There are other sets called "fully charged", which has a subset "positive"; and also there are the aliphatic and aromatic sets. The single letter IUPAC nomenclature for amino acids is used (Eidhammer, Jonassen and Taylor, 2004).

Colour schemes:

The colour schemes of the different amino acids, helps to point out the conserved residues or regions. There are different colours in this colour scheme, and this is represented in Table 3.1.

Table 3.1 Colouring schemes in the MSA

AMINO ACIDS	COLOURS	PROPERTIES
AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RHK	MAGENTA	Basic
STYHCNGQ	GREEN	Hydroxyl + Amine + Basic - Q
Others	Gray	

Table 3.1 shows the colouring scheme of the 20 naturally occurring amino acids. Table 3.1 adapted from (http://www.ebi.ac.uk/clustalw/#)

Table 3.2 Colouring scheme in a MSA according to physiochemical properties

Table 3.2 shows the colouring scheme according to the physio-chemical properties in Genedoc. Figure adapted from the Genedoc Multiple sequence alignment (http://www.psc.edu/biomed/genedoc/tutorial.htm).

A part of the MSA is shown below as an example to show the symbol and colour schemes as used in Figure 1.4.

Figure 3.2 Example of an MSA showing different conservation notation

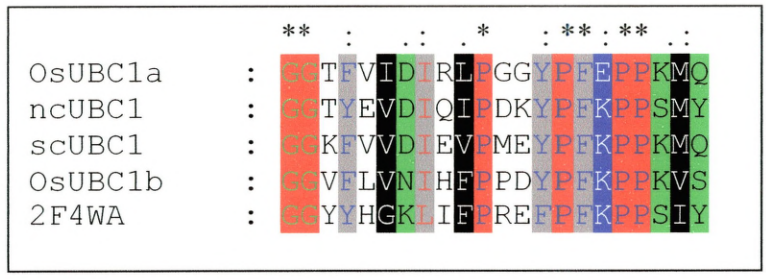


*Figure 3.2 illustrates an example of a MSA showing different conservation notations. Key to symbols: * = which means completely conserved; : = which means highly conserved; and . = which means less conserved (Poirot et al., 2003).*

Multiple sequence alignment can also help to identify the positions in homologous sequences that have descended from a common ancestor. In general the more diverged the sequences are to each other, the more difficult it is to align them. As multiple sequence alignment is the first step before phylogenetic and functional analysis, its errors will have broad effects on further analysis down the line. For example, different phylogenetic results can be obtained from different sets of multiple sequence alignments.

The accuracy of a multiple sequence alignment depends on the percentage of identical nucleotides or peptides in sequences. Generally it is known that if the sequence identity is more than 80%, then the aligned sites are true homologues, but as percentage identity decreases below 65% the probability of sequences to be correctly aligned also decreases. When the percentage of identical sequences further drops to below 50%, then the alignment becomes less accurate. The MSA algorithms set, also allow the insertion of gaps, so that identical nucleotides or peptides align between sequences (see Figure 3.3). A gap occurs in an alignment when either of the contributing sequences is completely lacking an amino acid at an alignment position (Altschul et al., 1997).

Figure 3.3 Illustration of the use of gaps in MSA

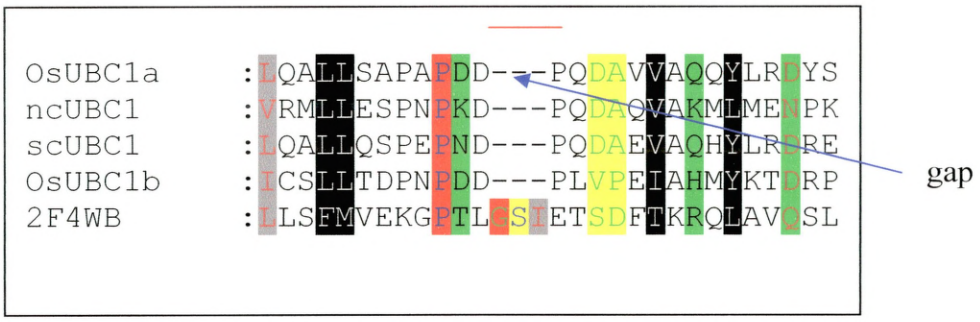


Figure 3.3 is a multiple sequence alignment of few UBCs illustrating how gaps are introduced in order to correctly align 2F4WB (the pdb ID of UBE2J2) with the other sequences. The line in red above the alignment is the gap in the multiple sequence alignment. Gaps are represented by a dash or a series of dashes.

3.1.3. Scoring matrices in sequence alignment

The scoring matrices are used internally by all sequence alignment programs. There are many different types of scoring matrices that have been deviced for sequence alignments of peptide sequences. A scoring matrix is generally a tool to measure the relationships between sequences. Through a succession of independent point mutation proteins evolve. Different mutation data matrices are appropriate for different alignments, and sometimes there are problems which are handled by editing manually (http://www.lmb.uni-muenchen.de/groups/bioinformatics/04/ch_04_3.html).
There are two different substitution scoring matrices the Dayhoff PAM matrix and the Henikoffs BLOSUM matrix.

3.1.3.1. DAYHOFF PAM matrices:

PAM called as point accepted mutation is a widely used scoring matrix, was developed by Dayhoff and coworkers (Dayhoff, Schwartz and Orcutt, 1978). Multiple sequence alignments were used for scoring amino acid pairs, by examining global alignments of closely similar sequences, and observing the frequency with which amino acid substitutions occurred. 1PAM represents an evolutionary distance in which 1% of the amino acids have been changed. There are a series of PAM matrices, which are appropriate for use with different sets of sequences separated by different evolutionary

distances. The different PAM matrices are PAM250, PAM120, PAM100, and PAM40. Among these PAM250 is the most widely used scoring matrix.

PAM250 gave a consistently higher significant score than any other PAM matrix. It is recommended to be used for pairwise comparison of proteins, which are known to be distantly related (Schwartz and Dayhoff, 1978).

PAM200 matrix is considered appropriate, when the sequences to be compared are thought to be related (Altschul, 1991).

PAM120 is considered to be the best option in a situation where the relationship between the sequences is not known (Altschul, 1991).

3.1.3.2. HENIKOFF'S BLOSUM matrix

The BLOSUM (Block Substitution Matrix) was developed by Henikoff and Henikoff (Henikoff and Henikoff, 1992). The difference between the BLOSUM scoring matrix and the PAM matrix is that, Henikoff took the approach of analysing multiple local alignments of distantly related sequences. Hence the substitution was considered are those amino acids occurring in relatively conserved regions of the protein rather than over the entire length of the sequence. PAM matrix was developed on the basis of global alignment of very similar sequences. There are different types of BLOSUM matrices, for example BLOSUM90, BLOSUM80, BLOSUM62, and BLOSUM30. These different matrices were obtained by varying the clustering threshold, for example, the BLOSUM80 was derived by using a threshold of 80% identity. Local alignments of related sequences were used to generate a database. Within this alignment, the sequences were clustered into groups, where the sequences are similar at a certain threshold value of percentage identity. Each of these groups is then used to calculate substitution frequencies for all pairs of amino acids. This is used to calculate a BLOSUM matrix. Thus the higher BLOSUM matrices are suitable for less divergent sequences, and the lower numbered BLOSUM matrices are suitable for sequences that are highly evolutionary divergent.

When comparing the two matrices (PAM and BLOSUM), it is seen that PAM250 is comparable to BLOSUM45, while PAM120 is comparable with BLOSUM80. BLOSUM62 is an intermediate in both clustering percentage, and is comparable to PAM160. It is also seen that BLOSUM62 is tolerant to substitution involving hydrophobic amino acids, but not to hydrophilic amino acids. BLOSUM 62 is typically

more tolerant to mismatches than any other matrices, especially for rare amino acids like cysteine and tryptophan. Hence BLOSUM62 is generally taken as a default scoring matrix (Henikoff and Henikoff, 1992; Barton, 1996).

It could be summarised that no single matrix is perfect for all sequence comparison. Substitution tables derived from structure comparison, will probably give the most reliable data, as more and more three dimensional structures are determined (Eidhammer, Jonassen and Taylor, 2004).

There is a high probability of incorrect inference of identity by an algorithm when the true variation among sites is very large. It may be possible to increase the percentage identity of sequences, by altering gaps or mismatch penalties, but these changes might alter the accuracy of the multiple sequence alignment. Though the sequence length does not affect the mean alignment accuracy, it affects the accuracy of evolutionary distance estimates. It is also known that the sequence length and the effect of alignment accuracy have no interaction on evolutionary distance estimation. Distance estimates are partially better with longer sequence that with short ones.

So it can be said that increased identity among sequences, helps to increase the accuracy of a multiple sequence alignment. Among all alignment programs, the most commonly used is the ClustalW, where it implements the most commonly used pair wise alignment method, the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). Hence it can be said that the use of aligned sequences, as for example, the identification of conserved sites, are highly dependent on the accuracy of the alignment. So, even a small error in it will have a large effect on the result. Evolutionary distance estimation is also somewhat robust to errors of multiple sequence alignment of moderate divergence (greater than 50% identity) (Rosenberg, 2005).

The BLOSUM62 scoring matrix, is designed by taking a big database of trusted alignments. The pairwise sequence identity is counted and a threshold value is taken into consideration of 62%. So all sequence identities falling around 62 % or less were considered in the BLOSUM62 target frequency; those identities falling in the higher region of 80% were considered in the BLOSUM80 frequency range. Similarly all sequence identities that are in the region of around 45%, were considered in the BLOSUM45 frequency range.

For calculating a score s(a,b) for aligning two residues a and b, the equation is as follows:

$$s(a,b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$

The likelihood that the two residue a and b are correlated because they are homologous, is indicated by the numerator in the equation (p_{ab}), which is the target frequency (the probability that is expected of the two residues a and b aligned in homologous sequence alignment). The denominator of the equation $f_a f_b$ is the likelihood of a null hypothesis, which indicated that these two residues a and b occurs independently and are uncorrelated and unrelated. Hence fa and fb are considered as the background frequencies, which is the probability that is expected to observe on an average amino acids a and b in any protein sequence. λ is the scaling factor, which allows rounding off all terms in the scoring matrix, to sensible integers.

If residues a and b is expected to be aligned together in homologous sequences more often rather than occur by chance, then the score is positive otherwise negative (Eddy, 2004).

3.2 Phylogeny

Molecular phylogeny is based on the comparison of DNA or protein sequences, and is used in estimating the evolutionary past. Phylogeny is a major tool and seems to find a very broad application in research from different backgrounds (Whelan, Li and Goldman, 2001).

A phylogenetic tree is composed of branches, which connect to nodes and from which it diverges into further branches. These branches and nodes can be internal or external. The first node from where several branches come out is called the internal node. This internal node could be called the last common ancestor (LCA) as everything arises from it. The nodes that are at the later stages of the phylogenetic tree is known as the external or the terminal node.

Figure 3.4 External node and internal node of a Phylogenetic tree

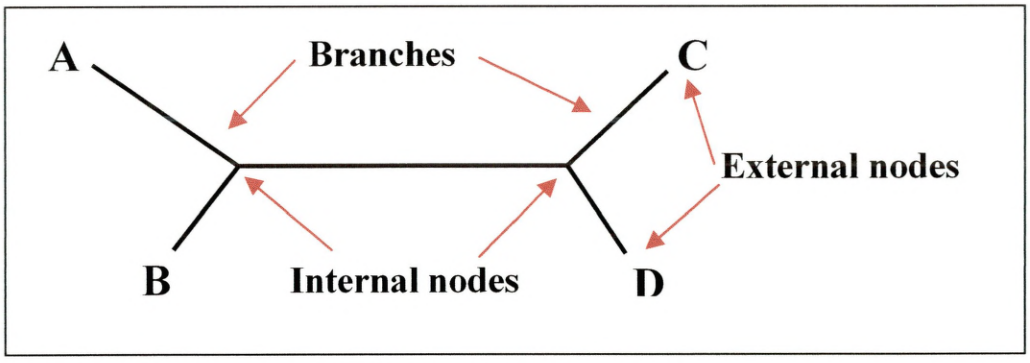


Figure 3.4 is an unrooted phylogenetic tree, showing the internal and external nodes. The figure was adapted from (http://trishul.sci.gu.edu.au/courses/ss13bmm/phylogeny_partA.doc).

The phylogenetic analysis consists of building up of a tree of multigene families where the internal nodes correspond to the gene duplication event, or in the case of a single gene from different species, the internal nodes corresponds to the speciation event, or a combination of both may be represented by the node. There are 3 types of groupings of a phylogenetic tree, the monophyletic, paraphyletic and the polyphyletic groups. When a group arises from a unique common ancestor, it is called a monophyletic group. When a group excludes some of its descendents, it is called a paraphyletic tree. When a group consists of members that superficially resemble each other, or have similar ancestral relationships, it is called as a polyphyletic group (Baldauf, 2003).

Figure 3.5 Groupings of a phylogenetic tree

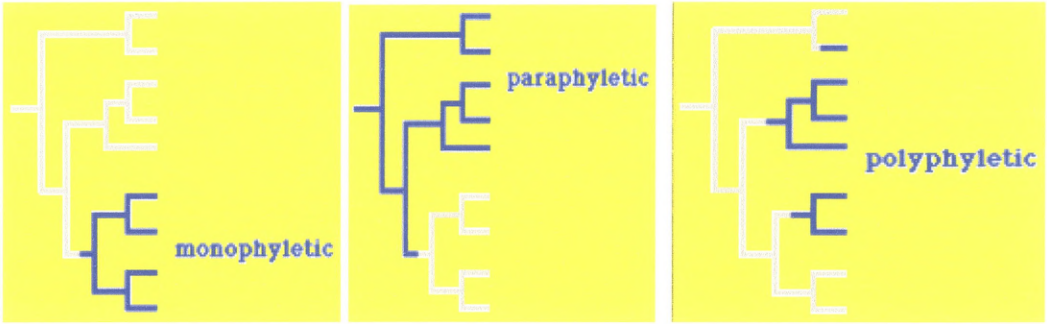


Figure 3.5 shows the three different groupings of a phylogenetic tree which are Monophyletic, Polyphyletic and Paraphyletic. (Figure 3.5 adapted from http://tolweb.org/tree/home.pages/glossary.html#polyphyletic).

A phylogenetic tree grows and as it becomes complex, the nodes expand and turns the tree on to its side, where the labels are then read horizontally, which makes it easily read and annotated. All branches can rotate from the plane of their nodes, and are usually drawn with proportional branch lengths. The length of a branch between two nodes corresponds to the evolutionary time, so the longer the branch, the more relatively divergent are the sequences attached to them. Another alternative method of drawing a phylogenetic tree, is to display the branching patterns only (cladograms), where the length of branches have no meaning, but it is rarely used.

The phylogenetic tree has a base, called as a "root", which is the oldest point of the tree. The root of a phylogenetic tree implies the order of branching that is, which sequence shares common ancestors with whom. A tree can only be rooted by keeping an external point as a reference called as an outgroup which is not a common member to the group of interest. There are two types of sequence homologues, orthologues and paralogues. Orthologues are genes of different species that have evolved from common ancestors maintaining similar function, whereas paralogues are genes with common ancestors that have undergone duplication resulting in a different, though related function (Nagl, 2003). Gene duplication is the prime factor to be considered when discussing orthologues and paralogues. Represented in Figure 3.6 is an example to illustrate gene duplication.

Figure 3.6 Gene duplication

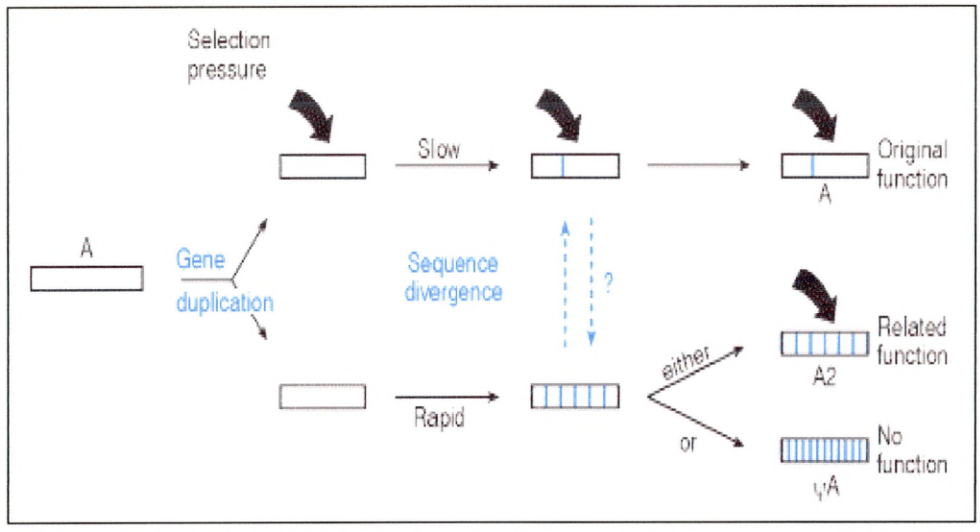


Figure 3.6 illustrates that gene duplication can lead to either acquiring a novel function or could form a pseudogene. Figure adapted from Strachan & Reid 1999. The example shown here is the duplication of gene A, which results into two gene copies. In the presence of selection pressure the gene copy (upper figure) has maintained the original function. Whereas the other gene copy (bottom) will still continue to be expressed but in the absence of selection pressure will mutate (represented by vertical dotted lines). This mutation could be deleterious and could form a non-functional pseudogene, whereas in other cases it may alter the function, which is selectively advantageous (represented A2).

3.2.1 Tree-puzzle

During the course of evolution, most nucleotide changes will occur in non-coding regions, as changes in the coding region will often be deleterious. In addition changes to regulatory or intronic splicing regions may also be deleterious as they may affect gene expression or correct splicing of genes. Therefore, what commonly occurs in many parts of the genome is heterogeneity in evolutionary rate. The phenomenon of rate heterogeneity is so common that it is assumed that the more conserved the sequences are, the more functional it is. Sequence that have changed rapidly in the course of evolution, have probably lost or have no function (e.g. pseudogenes). In a protein, some of its region evolve more rapidly than others, and is the same with different chromosomal regions, evolving at different rates (Schmidt, Strimmer and Haeseler, 2004). It is also known that the whole segment of the genome may evolve at different intrinsic rates. In carrying out a phylogenetic analysis, the effect of rate variation is estimated. The different parameters for estimating the rate variation are 1) "gamma" distributed substitution rates, 2) discretely distributed rates of variation, and 3) Hidden Markov Models (HMM) (Yang, 1994; Gu, Fu and Li, 1995). These tests assign different rates at different sites of a sequence, but they do not test for the significance of these differences. The reason for this is that individual sites can have different rates, whereas a single site is just one realization, which cannot be easily analysed. Hidden Markov model determines if the likelihood of variation is higher under a particular rate scheme (Hartmann and Golding, 1998).

In phylogenetic analysis the first criteria after the MSA is to estimate the coefficient of variation (CV), and in order to do that it is required to know the parameter "α". α is the index of the degree of among site rate variation (Gu and Zhang, 1997). The gamma distribution is used as a model for the rate variation among sites. The alpha parameter of gamma distribution describes degree of variation of rates across position. The coefficient of variation (CV) is calculated as $1/\sqrt{\alpha}$ (where α is the gamma distribution). Changing the α parameter of the rate variation, has an effect on the alignment accuracy. A small α value suggests that rates differ significantly over sites, while a very large α value means roughly equal rates. The greater the value of α is, the weaker is the rate variation, and as $\alpha = \infty$, it implies a uniform rate among site (Gu and Zhang, 1997; Yang, 1993; 1996). There are two main methods of calculating the value of α from the peptide sequences; these are a) Maximum Likelihood and b) Maximum Parsimony. The

Maximum likelihood method is very time consuming and very difficult to apply for more than five sequences. The second method which is the parsimony method, is computationally faster and is most used (Gu and Zhang, 1997; Yang, 1996; Goldman and Whelan, 2000). There are several different rate models e.g. a) Uniform rate model, b) discrete gamma distribution rate model, c) mixed model (gamma plus invariable sites) (Lio and Goldman, 1998; Zhang and Gu, 1998; Yang, 1996).

3.2.2 Phylip 3.6

Phylip is a phylogenetic package created by Joseph Felsenstein at the University of Washington (http://evolution.genetics.washington.edu/phylip.html). Phylip can be run on the Unix platform, but windows version is also available. This program reads the MSA which is saved only in the phylip format (.phy). The different phylip programs run in a sequential manner, where the output from the first program is used as the input for the next program. This package has phylogenetic analysis methods such as maximum likelihood, distance matrix (neighbour-joining, Fitch), maximum parsimony, and statistical estimation tool Bootstrapping.

Maximum likelihood program in phylogenetic analysis, is the statistically most rigorous algorithm, but has the major disadvantage, that it is time consuming, and it only accepts 50 sequences to do the analysis at a time. Neighbor-joining is the other most preferred phylip program which is used here to do the analysis (Felsenstein, 2001; Kuhner and Felsenstein, 1994).

3.2.3 Distance matrix method (Neighbor-joining)

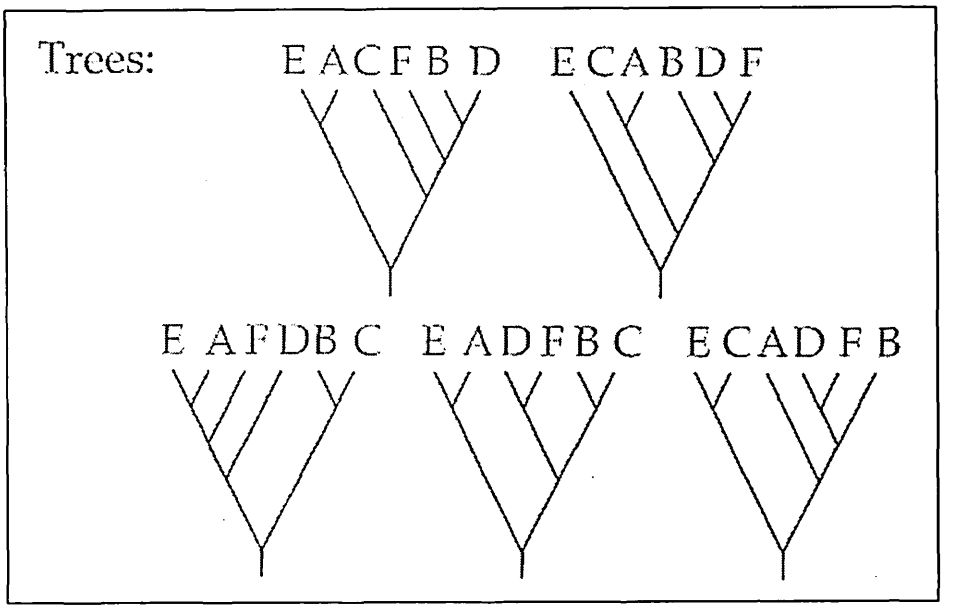
Saitou and Nei (1987) designed the Neighbor-Joining (NJ) algorithm, which works by clustering. The advantage of the NJ algorithm is that it is fast and fairly accurate. (Felsenstein, 2004). In the neighbour joining method, the phylogenetic tree is reconstructed, and the length of the branches of the tree is computed. The nearest nodes of the tree are chosen, and are defined as neighbors in the tree. This is done consecutively, until all the nodes of the tree are paired together (http://www.tau.ac.il/~doronadi/neighbourJoining.pdf#search=%22Neighbour-joining%20%22).

3.2.4 Bootstrapping

Bootstrapping was invented by Bradley Efron (1979) as a general purpose statistical tool and is frequently used to access the uncertainty of estimates of the phylogeny. A bootstrap is the statistical technique for estimating the variability of an estimate. It involves resampling from one's sample with replacement, and making an assumed sample of the same size. It involves resampling points from ones own data, with replacement, to create a series of bootstrap values of the same size as the original data. In phylogeny the proper method of resampling is to keep all of the original species, and the sampling is done with replacement on the assumption that the characters have evolved independently. A phylogenetic tree is constructed showing all the inferred monophyletic groups that occurred in a majority of the bootstrap samples. It has been shown that if a group in the particular branch appears 95% of the time, then it is taken to be statistically significant (Felsenstein, 1985). In bootstrap sampling there are several trees with different branching, and to keep track of all is a tedious task. Hence to get a consensus of all the trees there is the program called *majority rule consensus* tree, which summarises all the trees and plots a single tree, as illustrated in Figure 3.7.

Figure 3.7 Consensus in phylogeny

A. Five different phylogenetic trees with species appearing in different branches



B. Number of times each partition of species found

AE | BCDF	3
ACE | BDF	3
ACEF | BD	1
AC | BDEF	1
AEF | BCD	1
ADEF | BC	2
ABDF | EC	1
ABCE | DF	3

C. Consensus of the five different trees

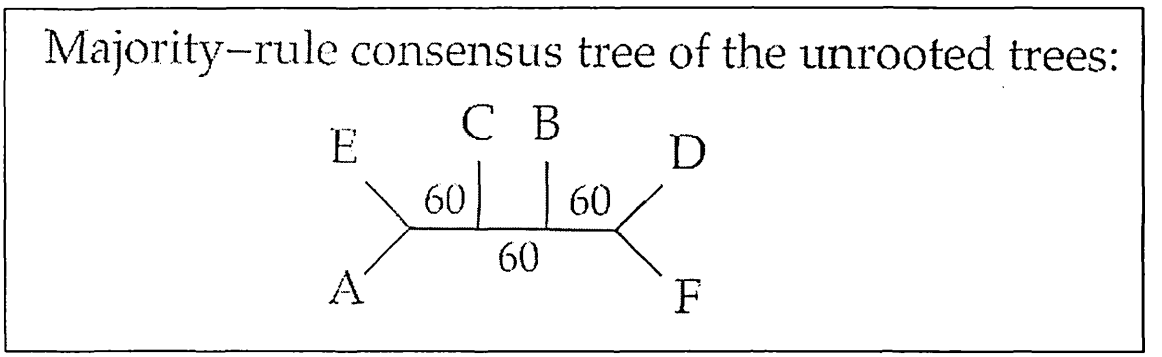


Figure 3.7 shows trees obtained in the phylogenetic analysis, where all species appear in separate branches in each trees as can be seen in part A of the figure. Part B of the figure shows the number of times each species appearing in the branch cluster in the five different trees represented above. Part C of the figure shows the consensus unrooted phylogenetic tree by the Majority-rule consensus method. The consensus tree has the percentage of support for each interior branch shown. Figure 3.7 adapted from (Felsenstein, 2004).

A consensus tree is a tree that summarises a series of trees. There are three different types of consensus trees:

1) Strict consensus

2) Majority rule consensus

3) Adams consensus.

Margush and McMorris's *Majority rule consensus* tree is simply a tree that consists of those groups that occur in a majority of the trees, and it is this consensus method that is mostly used (Felsenstein, 2004). The majority rule consensus tree is found by tabulating all groups that occur on all trees and retaining those that occur on a majority of the trees. The result is a single tree when it is used on the bootstrap estimates. All of the groups that appear on it are present in more than 50% of the bootstrap estimates. The values at the nodes of the tree in Figure 3.7C, are the bootstrap value, indicating the percentage of times the groups appear together in the branch (Felsenstein, 2004).

3.2.5 Phylip tree plotting programs: Drawgram and Drawtree

Drawgram and drawtree are tree plotting programs, that read the tree description in a file, and plots the phylogenetic tree. Drawgram plots a rooted tree called Cladogram and Phenogram, whereas drawtree plots an unrooted tree. Drawgram and drawtree plots the tree which can be saved in a postscript format, where this postscript file is taken to the other drawing package called Adobe Illustrator where other changes, such as writing the bootstrap values to the individual branches, altering the font size of the names, can be made to get the well represented colourful tree. (http://evolution.genetics.washington.edu/phylip/doc/draw.html)

TreeView is another tool used to draw Phylogenetic trees, but are not as portable as DRAWGRAM and DRAWTREE.

3.2.6. Summary of the advantages and disadvantages of Phylogenetic programs

The preceding sections introduced a range of readily available and widely used software tools and analytical procedures for the preparation of phylogenetic trees from multiple sequence alignments (MSA). The following Tables 3.3.1 & 3.3.2 summarises the main advantages and disadvantages of the most commonly used analytical procedures.

Table 3.3.1 Main advantages of the three phylogenetic programs

	Maximum Likelihood	Neighbor-joining	Maximum parsimony
1	consistent	consistent	Rapid
2	Most accurate	Rapid and analyses maximum number of sequences	Logical model
3	Gives robust tree	Reconstructs unrooted tree	Most uncomplicated model
4	-	Responsive to bootstrap analysis for very large trees.	-

Table 3.3.2 Main disadvantages of the three phylogeny programs

	Maximum Likelihood	Neighbor-joining	Maximum parsimony
1	Very slow	Does not give robust tree	Non-consistent
2	Can take only small datasets	Accuracy declines if datasets have finite sequences	Accuracy level of giving true tree is quite low.
3	Does not give unrooted tree	-	Discards large amount of data as they lack informative sites.

CHAPTER FOUR

INTRODUCTION OF CONSURF ANALYSIS

4.1 Orthologous sequence and conservation of structure

Consurf analysis is used to identify functionally important regions of a protein of known three dimensional structure. It estimates the degree of conservation of the amino acid sites by comparing it with its close homologues. It is an automated web based tool, where after identifying the functionally important regions, maps it on the surface of the protein (Glaser et al., 2003).

The analysis in Consurf is done automatically, where the only input is the query PDB ID and the multiple sequence alignment of its homologues in any one of the seven ClustalW format. This part is optional, as the Consurf server itself finds its homologues of the query protein, and does a multiple sequence alignment (MSA) in ClustalW and finds out the conservation sites along the whole length of the protein sequence. It then constructs a phylogenetic tree of the MSA using the neighbour-joining algorithm, and the position specific conservation scores are computed using either the Bayesian or the Maximum Likelihood algorithm. These conservation scores are divided into nine grades in different colours, where grade-1 indicates *most variable position* and is assigned the colour Turquoise, grade-5 which is the *intermediate conservation,* is given the colour white, and grade-9 which is the *most conserved* is assigned the colour maroon. These nine colour conservation grades are plotted on the three dimensional structure of the query protein (Landau et al., 2005; Armon, Graur and Ben-Tal, 2001).

CHAPTER FIVE

INTRODUCTION TO STRUCTURE PREDICTION BY X-RAY CRYSTALLOGTAPHY

5.1. Introduction to all crystallographic methods

5.1.1. Cloning

5.1.1.1. DNA cloning

DNA cloning is the selective amplification of a desired fragment of DNA molecules with restriction endonucleases and ligases. This is used to study its structure and function, by DNA sequencing, in-vitro expression studies, to name a few. There are two different DNA cloning approaches:

1) Cell-based DNA cloning,

2) Cell-free DNA cloning.

The cell based DNA cloning is carried out in-vivo, whereas the cell-free DNA cloning is carried out in-vitro. The cell free DNA cloning is enzyme mediated and the reaction involved is called as "Polymerase Chain reaction" (PCR). In cloning, plasmids act as "cloning vector". Plasmids are small circular double stranded DNA molecules. The plasmids provide the replicative ability to enable the cloned gene to be propagated inside the host cell. The plasmids have a point of replication, which is recognized by the DNA polymerases that normally replicate the bacterial chromosomes (Brown, 2002; Strachan and Read, 1999).

5.1.1.2. Polymerase chain reaction (PCR)

The rapid method of amplifying the target DNA sequences, present within a source of DNA, is PCR (Polymerase Chain Reaction). Two oligonucleotide primers (amplimers) are designed, which are specific for the target sequence. The primers are usually about 15-25 nucleotides long.

The steps in a PCR reaction are:

1) Denaturation (usually at a temperature of 93-95°C).

2) Reannealing (at a temperature between 50-70°C, depending on the Tm. The annealing temperature is usually about 5°C below the Tm),

3) DNA synthesis (usually at a temperature of about 70-75°C).

The template DNA is denatured first, and the primers are added to it, which binds specifically to complementary DNA sequences at the target site. The initiation of synthesis of new DNA strands is carried out by the presence of heat stable DNA

polymerase (Taq polymerase) and DNA precursors (dATP, dTTP, dCTP, dGTP). The newly synthesised DNA strands are complementary to the individual DNA strands of the target DNA segment, and which overlaps each other. The newly syntheised DNA strands acts as templates for further DNA synthesis. Hence PCR is a chain reaction and around 25-30 cycles are given to get around 10^5 copies of the specific target sequence. This result is visualized as a band when subjected to gel electrophoresis (Brown, 2002; Strachan and Read, 1999).

5.1.2. Protein purification

5.1.2.1. Gel permeation chromatography

There are several different methods of purifying a protein sample prior to its crystallization. These are a) Adsorption chromatography, b) Partition chromatography, c) Ion exchange chromatography, d) molecular exclusion chromatography, and e) affinity chromatography. The method used in the thesis is "molecular exclusion chromatography. The parameter that is taken into account in gel permeation chromatography is the molecular size of the protein. The parameter of elution of the neutral lipids from the lyophilic dextran gels, is generally in the decreasing order of molecular weights, but there is another factor that affect this pattern, and that is the polarity of the lipid or the eluent (Yip, 1997; Calderon and Baumann, 1970).
Figure 5.1 next page illustrates gel permeation chromatography.

Figure 5.1 Gel permeation chromatography

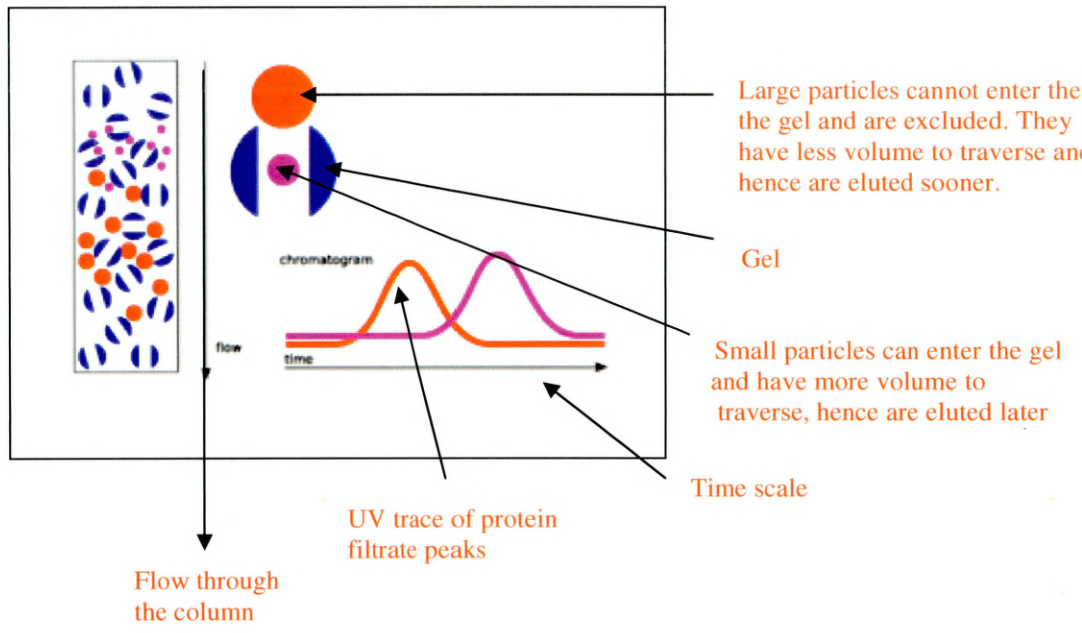


Figure 5.1 illustrates the principle of size exclusion chromatography, wherein larger molecules are eluted faster than the smaller molecules (Figure adapted from http://www.answers.com/topic/size-exclusion-chromatography).

The matrix of the molecular sieve chromatography is composed of highly cross-linked poly-dextrans e.g. Sephadex or any other similar material like polyacrylamide, which are in the form of microscopic beads. At the molecular level, these beads appear to be in the form of a sphere, which is perforated with a network of channels. In the process of molecular sieve chromatography, the protein solution gets fractionated according to the molecular size. The macromolecules penetrate the interior of the beads depending on its size. The macromolecular mixture is made to flow within the column consisting of the beads. Those macromolecules, which are of sufficiently small size, penetrate the beads. As these small size macromolecules have to spend more time in migrating about the interior of the beads, their speed in passing from the top to the bottom of the column is greatly reduced. The larger size molecules are excluded from entering the interior of the beads due to their bigger size and hence pass through the column faster. Hence various size molecules are eluted and fractions collected at different rates. Larger molecules will

appear first and successively low molecular weight components eluted later. The effective separation in this method is of macromolecules differing in molecular weight of not more than 10-20%. There are exceptions to this method as well as fibrous molecules, or a molecule in the shape of an ellipsoid with an extreme axial ratio, also behave as if it were of a higher molecular weight (McPherson, 1999).

Figure 5.2 Molecular weight selective curves for G-type Sephadex

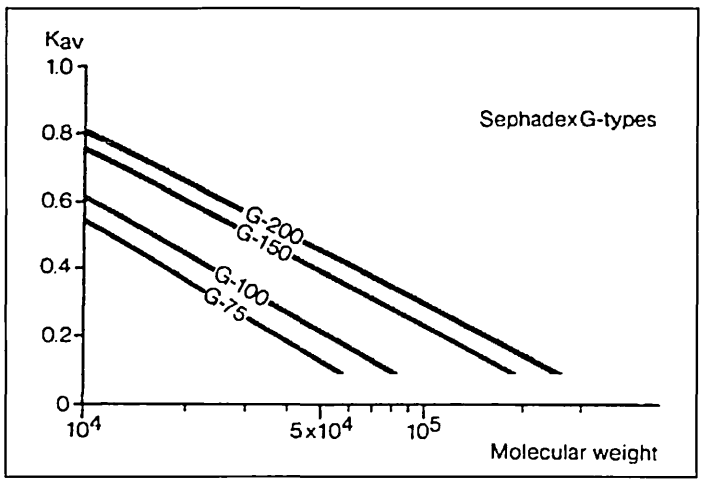


Figure 5.2 shows the molecular weight selective curves for G-type Sephadex, a popular gel filtration medium. On the X-axis is the protein molecular weight, and on the Y-axis is its migration speed (McPherson, 1999).

The advantage of gel permeation over ion-exchange chromatography is that gel permeation chromatography is gentler and faster as it does not require exposure to pH or ionic strengths and also does not require any gradient solution. In ion-exchange chromatography, a gradient is created of variable pH and ionic strength by mixing different buffers, where a compound is separated according to the degree of its ionic charge. In gel permeation chromatography, protein separation is according to its size, where the larger molecules are eluted first (McPherson, 1999).

5.1.3. Crystallography in general

In the early twentieth century Lawrence Bragg used the diffraction patterns from Zinc Sulphide, Sodium Chloride and Potassium Chloride to suggest the three dimensional structure of their ions in arrangements of their crystals. The principles of crystallography had been well understood since then. Myoglobin and Hemoglobin were

the first three dimensional protein structures determined by M.F. Perutz and J.C. Kendrew in the early 1960's (PDB codes 1MBN and 2DHB). Protein crystals include a large number of protein molecules arranged in a periodic manner. As crystals amplify the diffraction signal to a measurable level, this periodic arrangement is very essential, which would otherwise be weak if only a few protein molecules are available for analysis. The three dimensional structure of a protein molecule is built from an electron density map, which requires the measurement and fourier synthesis of amplitudes and phases from the reflections. Amplitudes are proportional to the square root of diffraction intensities, and are directly obtained from the diffraction measurements. The result of a diffraction experiment is obtained as an array of spots (reflections) that measures the intensities of waves scattered by the crystal. However during the diffraction experiment, the phase information is lost. Subsequently, there are two processes in X-ray crystallographic studies. Initial diffraction experiments are carried out to measure intensities (and ultimately amplitudes), and another set of experiments to determine the phases (Clegg, 1998; McPherson, 2003).

5.1.3.1. Properties of protein crystals

One or all of the following factors may influence size, quality, and X-ray diffraction characteristics of a protein:

(a) Air pockets

(b) Precipitant ions

(c) Impurities

(d) Ordered layers

(e) Disordered solvent

(f) Inclusions

(e) Covalent or non-covalently bound sugars, prosthetic groups, or other ligands

(f) Other, more complex layers

(g) Various defects and dislocations.

5.1.3.2 Methods of crystallisation (Vapour diffusion techniques)

This is the most important and successful method of crystallization, which produces very good diffraction quality crystals. The various vapour diffusion techniques are: Sitting drop, and hanging drop as follows:

Figure 5.3 Sitting drop vapour diffusion technique

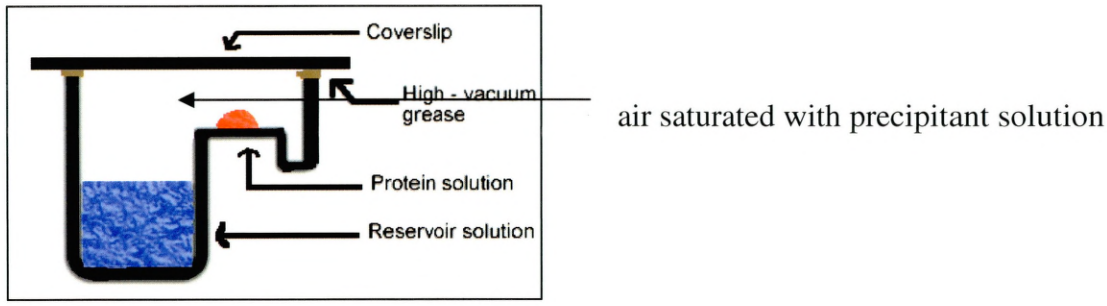


Figure 5.4 Hanging drop vapour diffusion technique

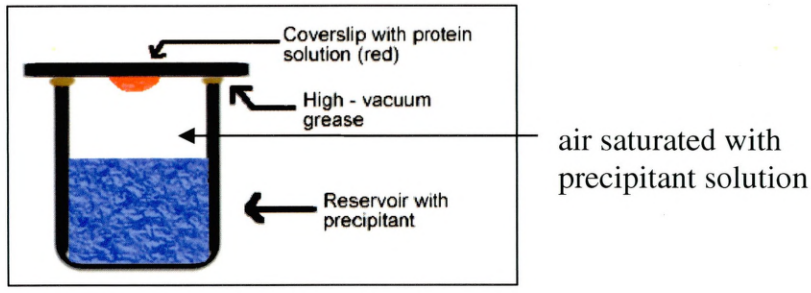


Figure 5.3 represents the sitting drop method, where the protein drop sits on a base above the reservoir solution, as opposed to hanging drop
(Figure 5.3 adapted from http://www.bio.davidson.edu/Courses/Molbio/MolStudents/spring2003/Kogoy/protein.html).

Figure 5.4 illustrates the hanging drop method. Reservoir solution (blue) usually contains buffer and precipitant. Protein solution (red) contains the same compounds, but in lower concentrations. The protein solution may also contain trace metals or ions necessary for precipitation of particular proteins. For instance, insulin is known to require trace amounts of zinc for crystallization (McRee, 1993; Rhodes, 1993).
(Figure 5.4 adapted from http://www.bio.davidson.edu/Courses/Molbio/MolStudents/spring2003/Kogoy/protein.html).

In the sitting drop vapour diffusion method wells are used of a great variety of plates available from the Hampton (http://www.hamptonresearch.com/assets/products/attachments/0000000001-0000000073.pdf). A droplet of 5-10 µl containing the macromolecule, a buffer and a precipitating agent is equilibriated against a reservoir of 1-25 ml, containing a solution of the same precipitant but with a higher concentration then that of the drop. Solvent equilibriation takes place, where the evaporation of the volatile species takes place until the vapour pressure of the droplet equals that of the reservoir, and crystals form in the droplet.

The hanging drop method is the same as of the sitting drop technique, with the exception that in this case the drop is inverted and remains intact by surface tension over the precipitant reservoir. Solvent equilibriation takes place as the sitting drop method, and crystals grow in the droplet.

The vapour diffusion techniques use plates with either 4 X 6 wells or for microbatch tests of 6 X 12 wells. It allows a wide range of choice of polymers, salts, and organic solvents with a high range of pH (see appendix). When a crystal is obtained in any particular condition, a second trial is set with that same condition in order to optimise the crystallization conditions, to produce X-ray diffraction quality crystals.

5.1.3.3. Crystal growth

In order to carry out X-ray crystallographic analysis, crystals of the protein must first be grown. Crystals are an ordered arrangement of molecules. When molecules are precipitated very slowly from supersaturated solutions, crystals are formed. "Hanging drop" and "sitting drop" are the two commonly used methods for making crystals. In these methods, the protein solutions are brought very gradually to supersaturation by vapour diffusion. The protein solution is mixed with the precipitant and a drop of this mixture is setup in a sealed atmosphere shared with a reservoir. Due to osmotic pressure the movement of water out from the protein drop to the reservoir occurs. The commonly used precipitants are generally salts (such as ammonium sulphate), polymers (for example, various molecular weight polyethylene glycols, PEGs) or organic solvents (2-methyl 2,4-pentadiol; MPD).

The three steps involved in crystal growth are as follows:

1) Nucleation- the formation of a "seed", which initiates crystal growth.
2) Growth- molecules of the protein in solution build up a crystal lattice around the initial seed.
3) Cessation of growth- The crystal stops growing when, either the system has reached an equilibrium, where the number of molecules joining the crystal lattice is the same as the number returning to solution; or the crystal redissolves due to factors such as the increasing osmolarity of the protein drop causing the equilibrium to reverse or the crystal to shatter. Crystals can also stop growing when external or damaged molecules are incorporated into the growing surface, resulting in defects that interrupt the established regular pattern. The figure below shows a plot of a theoretical phase diagram, showing the transition of a protein molecule from the soluble state to crystalline state (McPherson, 1999).

Figure 5.5 Crystallization phase diagram

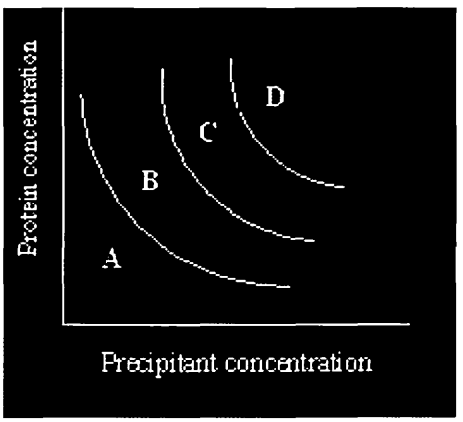


Figure 5.5 shows the crystallization phase diagram. (A) The undersaturated zone, where the protein remains soluble. (B) The metastable zone of supersaturation, where the protein is supersaturated enough to support crystal growth, but not high enough to initiate the nucleation process. (C) The nucleation zone, where protein is supersaturated enough to form crystal nuclei. (D) The precipitation zone in which the protein is highly supersaturated, leading to precipitation (Adapted from DeLucas et al., 2003).

As only well-ordered crystals of the target protein are capable of producing useful X-ray diffraction results, the design of a crystallization process is crucial. A typical

crystallization experiment starts with a screening process, which involves altering the physical and chemical environment of the target protein until some form of crystalline material is obtained. This is followed by an optimization process, which usually involves improving the size and diffraction quality of the crystal. Different crystals behave in different ways, where some may be stable for years, while others are not and provide only very limited opportunities to gather its diffraction pattern.

5.1.3.4 Factors that affect crystallization

Crystallisation is a complex process. For example many factors can affect crystallization, e.g. the pH of the protein drop, the precipitant type and concentration used, the temperature at which the experiment is carried out, and the addition of co-factors. A minute change in any of these parameters can make the difference between whether a crystal is formed or not. There is no particular way to determine which set of conditions will produce crystals for a particular protein. Therefore various potential conditions have to be tried in order to obtain a crystal. The purity of a protein sample is the most important factor for its crystallization as impurities may introduce imperfections, which are harmful for the growth of diffraction quality crystals.

Different degrees of supersaturation are created to determine the concentration of the protein required to produce X-ray quality crystals. Salts e.g. ammonium sulphate; high molecular weight straight-chain polymers, e.g. Polyethylene glycols (PEGs), and 2-methyl-2,4 pentane diol (MPD); and organic solvents like ethanol are some of the chemical precipitants used to achieve supersaturation of molecules.

Nowadays pentaerythritol propoxylate and pentaerythritol ethoxylate are used as newer precipitants (Gulick et al., 2002). The most successful salt reported in crystallization experiment is sodium malonate, which involved 23 proteins and compared 12 salts including ammonium sulphate, the conventional salt used for crystallization. Malonate is an organic acid with two negative charges at neutral pH. The role of these precipitants is to compete for water with proteins leading to intermolecular interactions that favoured the formation of crystals. pH of the protein solution is also an important factor for its crystallization as it affects the net charge on the protein molecule and the charged state of the exposed amino acids. Where a protein is less soluble, adjustment of its pH can lead to the formation of crystals. In most cases crystallization occur over a narrow range (<1 pH unit), but in some cases it occurs over a broad range of pH also. Usually

the pH around which protein crystallization occurs is the Isoelectric point of the protein, where the net charge on the protein is zero.

Temperature has an important role on protein solubility and the behaviour of the other components discussed before. The rate of equilibration is generally slower at lower temperatures. Temperature ranges from 0-4°C and 20-24°C are suitable to obtain most protein crystals. There are exceptions where a few proteins have crystallized at temperatures lower than 0°C and above 40°C. In recent times crystallisation is carried out in a robot where multi-well plates are set, with each well containing a different condition.

5.1.3.5 Crystal mounting for x-ray data collection

In order to obtain the X-ray diffraction pattern of the protein crystal, the crystal must be removed from the drop and placed in the path of the X-ray beam. Generally a loop of nylon is used to "fish" out the crystal from the protein drop. Once the crystal is in the loop, the base is attached magnetically to the goniometer, which allows the crystal to be aligned and rotated along the path of the X-ray beam. Surface tension helps the crystal to be attached in the loop during this process. The lifetime of a crystal may range from a few minutes to several days under normal laboratory conditions. But when the crystal is exposed to high intensity X-ray, the life time of the crystal usually decreases drastically. In order to prevent this denaturation of the protein crystal, a process called as "Cryo-crystallization" is used (Stum et al., 2004).

5.1.3.6. Cryo-crystallography

In this process the temperature of the protein crystal is lowered by rapidly passing a stream of cold liquid nitrogen through the mounted crystal. Also the lowering of the temperature of the crystal causes the lowering of the thermal motion of the constituent atoms, and the X-ray diffraction pattern is improved both in intensity and resolution. To avoid the freezing of the water in the crystal and other solutions, a chemical that acts as a "cryoprotectant" is mixed with it before the crystal is mounted. This allows the water to freeze as a glass. Chemicals commonly used as cryoprotectants are Glycerol making up to a concentration of 10 – 30% of the final solution. Other chemicals used are low molecular weight PEGs, ethylene glycol and sodium malonate (Rose and Wang, 1997)

5.1.3.7. X-ray diffraction

X-rays are a form of electromagnetic radiation with a wavelength in the range of 10^{-9} m – 10^{-8} m. For protein crystallography, there are two main sources of X-rays, "in- house" systems and synchrotron. The in-house system is relatively smaller than the synchrotron system. The synchrotron system is a particular storage ring, which accelerates electrons around a circular track (which is called the beam line) in a high vacuum at close to the velocity of light. The electrons emit an intense continuous spectrum of X-rays. This is caused by using magnetic fields to steer the electrons around the ring. The in-house system produces X-rays by focusing a beam of electrons on a small area of metal. The design of the in-house system is a "rotating anode X-ray generator", where the electrons are focused on a small area of a rotating copper cylinder. X-ray produced from synchrotrons, are much more intense than that produced from the in-house source and they have much better optional properties. As there are very few synchrotrons in the world, the early experiments are carried out in-house to determine factors like the best cryoprotectant (Drenth, 1999; McPherson, 2003).

Diffraction is a phenomenon that occurs when a wave passes through a slit or a series of small openings, where the size of the slit is comparable to the wavelength. X-rays are used to create "diffraction patterns" from the protein crystals.

Figure 5.6 X-ray diffraction patterns

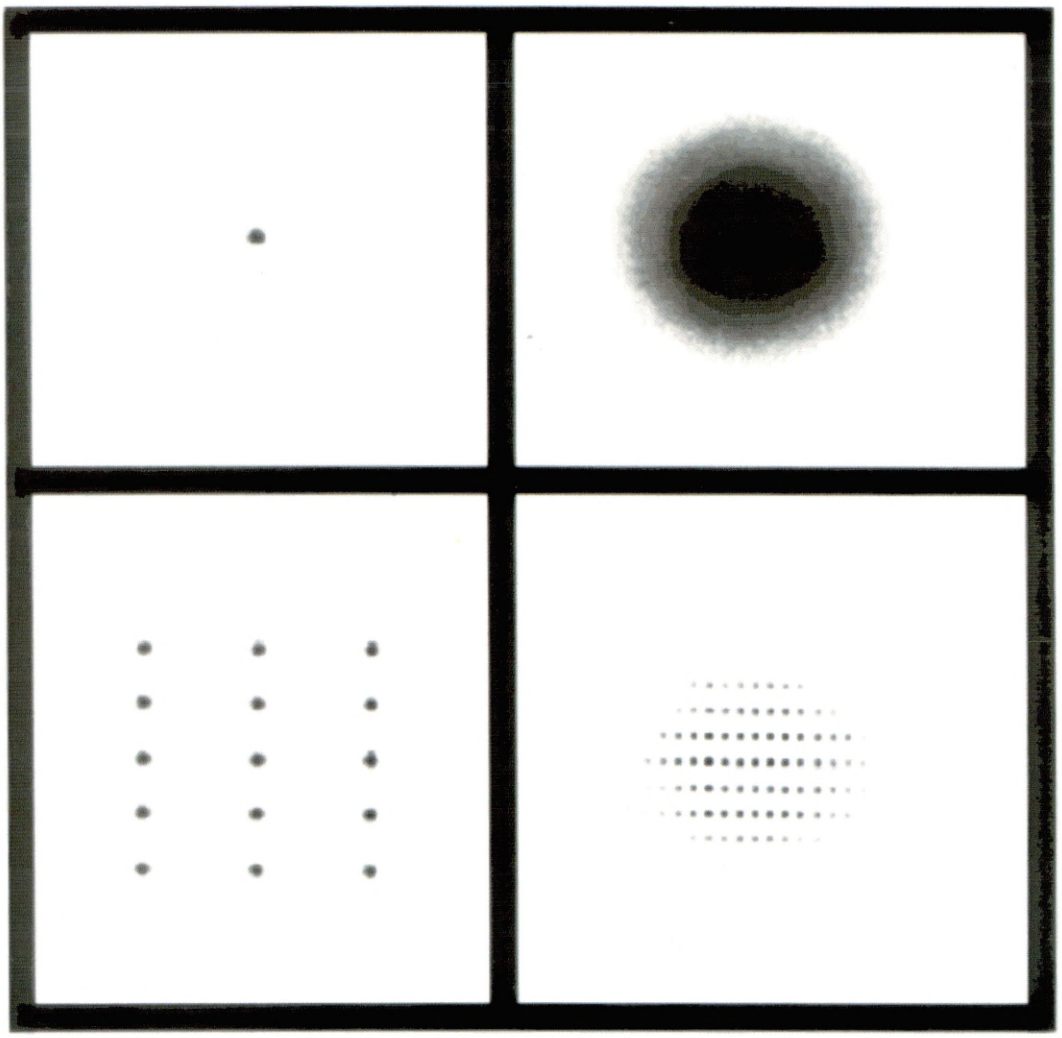


Figure 5.6 illustrate the patterns obtained by X-ray diffraction of an object. The sphere in the top left, gives the diffraction pattern shown in the top right. The bottom part of the figure depicts diffraction by a crystalline array of spheres, where to the bottom left is a cross section of the crystal, and its diffraction pattern to its right (figure adapted from Rhodes, 2000).

The lattice spacing of the crystal is smaller vertically (shown bottom left), and the diffraction spacing is smaller horizontally (shown bottom right). There is the inverse relationship between the spacing of unit cells in the crystalline lattice (real lattice) which is indicated in the bottom left part of the figure above, and the spacing of reflections in the lattice on the film (reciprocal lattice) as indicated in the bottom right (Rhodes, 2000; McPherson, 2003).

5.1.3.8. The arrangement of protein molecules in a crystal

From the diffraction pattern, the molecular arrangement of the diffracting object can be estimated. The diffraction pattern can be used to determine the shape of the protein and its arrangement within the crystal. A crystal is built up of a large number of identical units, called the "unit cells". The unit cells contain one or more "asymmetric units", the smallest unit of the crystal which can be used to generate the unit, and then in turn the complete crystal.

The crystal contains the unit cells in three dimensions, the lengths of three unique edges are **a, b,** and **c,** and three unique angles α, β, and γ illustrated in Figure 5.7.

Figure 5.7 **A triclinic unit cell**

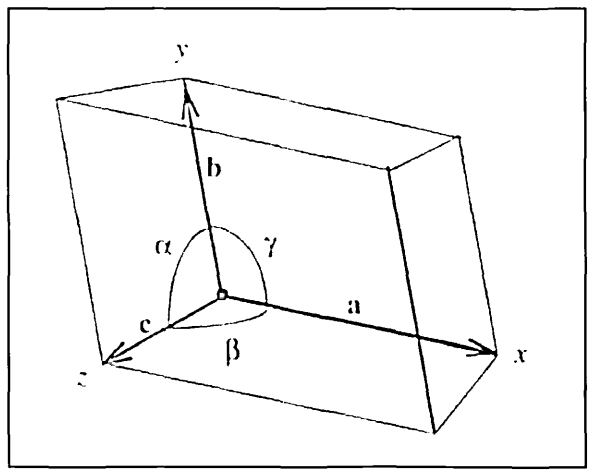


Figure 5.7 shows a general triclinic unit cell, with edges a, b, and c, and angles α, β, and γ (Figure adapted from Rhodes, 2000).

A crystal contains unit cells in three dimensions. The transformation of one unit cell into another is a pure translation, and is a form of symmetry. There are different types of

crystals with relationships between the cell edges and the angles between the edges, as given in Table 5.1.

Table 5.1 The crystal systems

Crystal system	Relationship between axes	Relationship between angles
Triclinic	$a \neq b \neq c$	$\alpha \neq \beta \neq \gamma$
Monoclinic	$a \neq b \neq c$	$\alpha = \gamma = 90°$; $\beta > 90°$
Orthorhombic	$a \neq b \neq c$	$\alpha = \beta = \gamma = 90°$
Tetragonal	$a = b \neq c$	$\alpha = \beta = \gamma = 90°$
Trigonal/rhombohedral	$a = b \neq c$	$\alpha = \gamma = 90°$; $\beta = 120°$
Hexagonal	$a = b \neq c$	$\alpha = \beta = 90°$; $\gamma = 120°$
Cubic	$a = b = c$	$\alpha = \beta = \gamma = 90°$

Table 5.1 has been created with reference to Rhodes, 2000; Blow, 2002 and McPherson, 2003.

It is important to determine the crystal symmetry, as the measurement of X-ray scattering by the crystal during the initial stages of the experiment can be determined. The symmetry of the crystal at the final stages of data interpretation is also significant, which would lead to the determination of crystal structure.

5.1.3.9 Diffraction of X-rays by crystals

Bragg's law

The diffraction pattern obtained from a crystal in a given orientation is determined by Bragg's law. A crystal is composed of many parallel mirrors, which is referred to as Bragg planes, which reflect X-rays. How rays scattered at the lattice points in one plane, all travel the same distance to the scattered wave front, is given in the first part of Bragg's explanation. Any plane of lattice points of a crystal acts as a mirror. Many parallel lattice planes are created by the crystal lattice. The three quantities that are dependent on for two planes to scatter in phase are:

-The wavelength λ of the X-rays,

-The spacing d between the planes,

-The glancing angle θ.

The planes are represented by a set of integers, h k l, referred to as lattice indices. According to the Bragg's law, when X-rays of wavelength λ comes in contact with a set

of parallel planes with the distance between the two planes d, at an angle θ, and are reflected at the same angle, diffraction occurs.

The equation is as:

$$n\lambda = 2d\sin\theta$$

(n is an integer, which denotes the path length between adjacent planes) (Palmer, 2001; Blow, 2002).

Figure 5.8 Diagramatic representation of Bragg's law

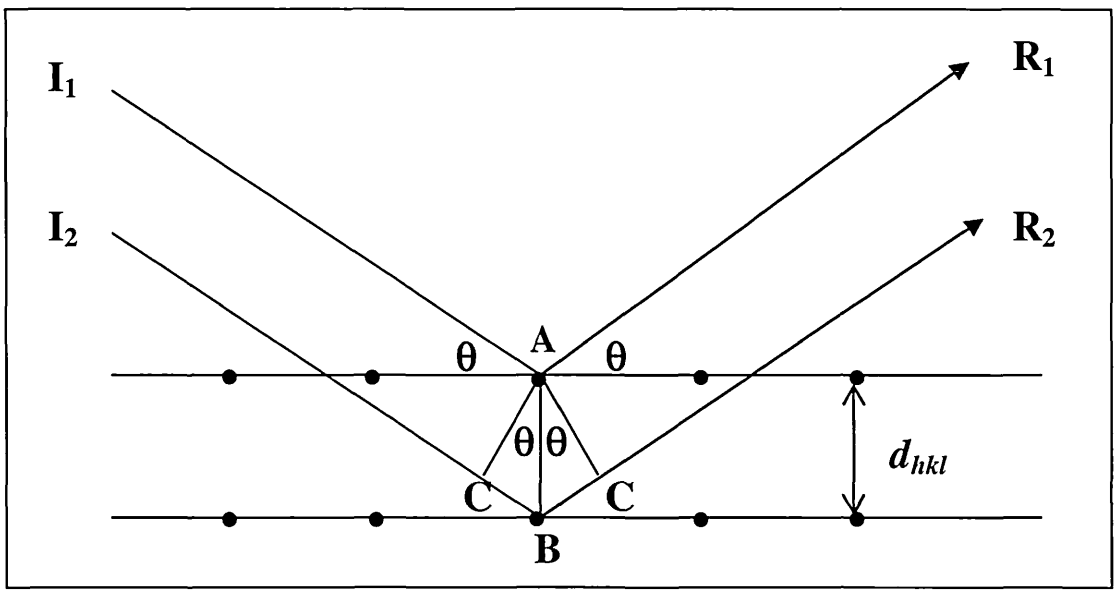


Figure 5.8 shows the conditions that produce diffracted rays according to Bragg's Law. The parallel horizontal lines represent two parallel planes with interplanar spacing d_{hkl}. The shaded circles on the horizontal lines represent atoms. The incident (I_1 and I_2) and reflected (R_1 and R_2) rays make an angle θ with the planes. Since two angles are equal, provided corresponding sides are perpendicular, the angle CAB also equals θ (figure adapted from Rhodes, 2000; Palmer, 2001).

5.2. Circular Dichroism spectroscopy

A very important tool for investigating the structural properties of a protein is the Circular Dichroism (CD) spectroscopy. The secondary structure content of a protein, its tertiary structure, and fold recognition can be analysed by CD spectroscopy. The reason why circular dichroism spectroscopy has become the method of choice is its relative speed of analysis, the small amount of sample required, and the information about the overall structure of a protein that can be obtained (Andersson, Carlsson and Freskgard, 2001).

In the recent time, a high number of resolved protein structures have been submitted to the protein data bank (PDB). X-ray crystallography or NMR, are the two important methods for structure prediction of a protein, but as they are very expensive and time consuming, circular dichroism spectroscopy has become increasingly recognized as a structural analysis technique. The main advantage in CD spectroscopy is that the analysis can be performed under the conditions, in which proteins actually operate, that is generally in solution (Kelly, Jess and Price, 2005).

5.2.1. Mechanism of CD spectroscopy

CD spectroscopy can be observed for optically active molecules or in the case of molecule-ligand binding or if at least one of the interacting species is optically active (Drake, 2001). Plane polarised light is composed of two circularly polarised components of equal magnitude, one rotating clockwise (right handed R) and the other counter clockwise (left handed L). The principle behind the circular dichroism is also based on the differential absorption of these two (L and R) components. If the L and R components are absorbed to equal extent when a ray of light is passed through a sample, the recombination of L and R would regenerate radiation polarised in the original plane. The resultant radiation would be said to possess elliptical polarization if the L and R are absorbed to different extents (Drake, 2001; http://www.newark.rutgers.edu/chemistry/grad/chem585/lecture1.html).

Figure 5.9 Principle of polarization of light in CD spectrometry

A. Ordinary light B. Right circularly polarised C. Left circularly polarised

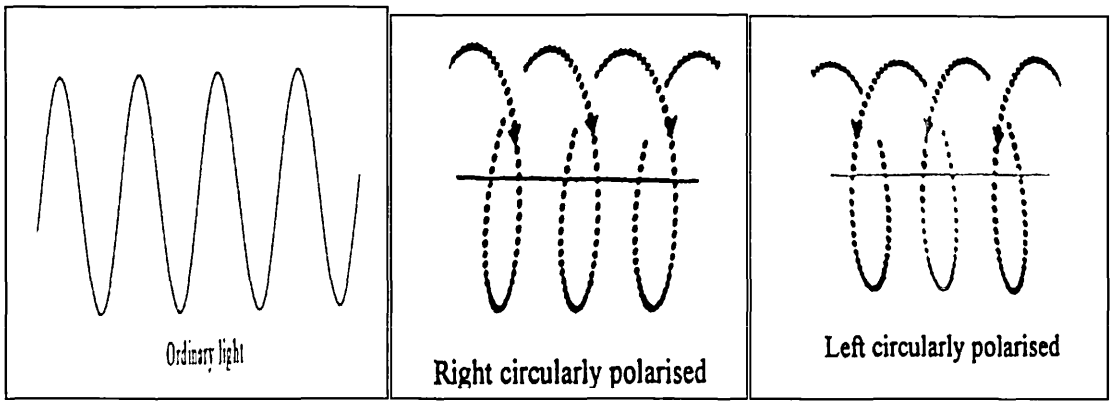


Figure 5.9 shows the ordinary light, the right circularly polarised and the left circularly polarised light in a CD spectroscopy (figure adapted from Drake, 2001).

When a chromophore is optically active, a CD signal is observed. The CD instrument measures the difference in absorbance between the L and R circularly polarised components ($\Delta A = A_L - A_R$), and is reported in terms of the ellipticity (θ) in degrees (Drake, 2001).

The far ultraviolet spectral range, ranges from about 190 – 250 nm, and it provides information on the secondary structural contents of proteins. Spectral analysis of below 190 nm provides further information about the protein conformations and folding. This is usually possible by using the CD spectrometers, using synchrotron radiation which can go to as low as 160 nm wavelength. (Sreerama and Woody, 2000; Wallace and Janes, 2001). The tertiary structure of a protein can be analysed from the near UV circular dichroism whose wavelength is in the range of 260 – 320 nm. The spectra in this region arises from the aromatic amino acids, like tryptophan shows a peak between 290 and 305 nm, tyrosine between 275 and 282 nm, and phenylalanine between 255 and 270 nm.

Figure 5.10 Far UV CD spectrum illustrating secondary structural features

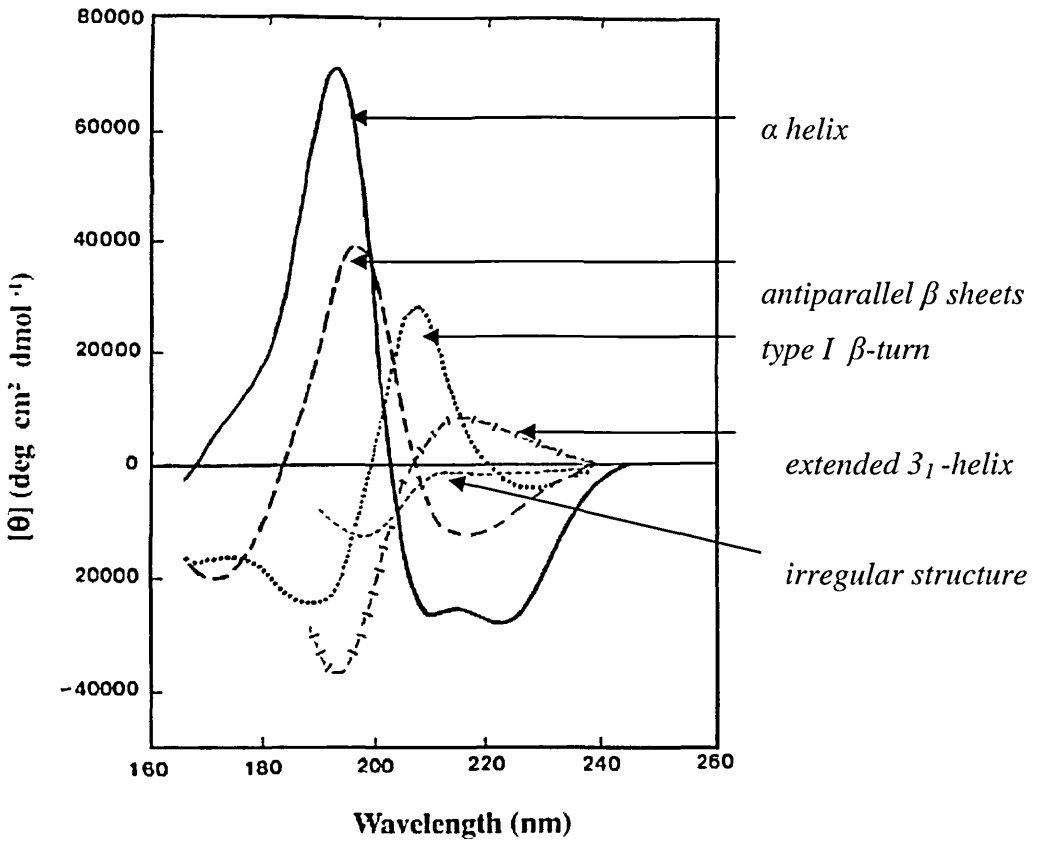
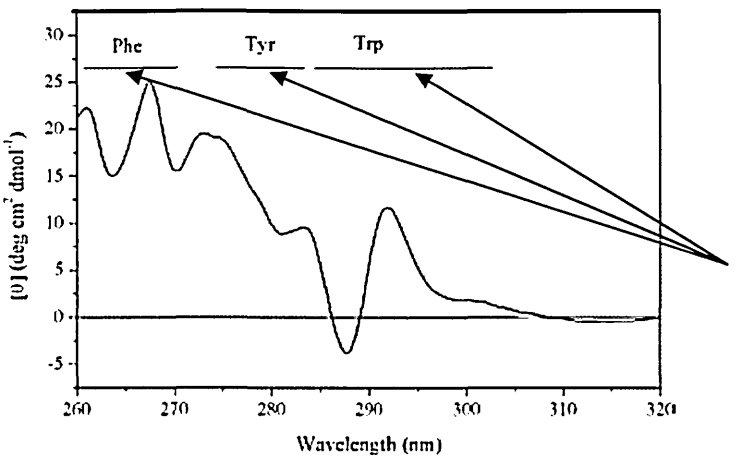


Figure 5.10 shows the far UV CD spectrum obtained in the range of 160 – 260 nm wavelength. The secondary structural features have been illustrated as marked in the figure (Kelly, Jess and Price, 2005).

Figure 5.11 Near UV CD spectrum arising from amino acid residues



The range of spectrum of the aromatic amino acids(Phe, Tyr & Trp)

Figure 5.11 shows the near UV CD spectrum obtained within the range of 260 – 320 nm wavelength of type II dehydroquinase from Streptomyces coelicolor as an example. The peaks represents the UV signals of the absorbance of Phenyl alanine (255-270 nm), Tyrosine (275-282 nm), and Tryptophan (290-305 nm) at different wavelength ranges (Kelly, Jess and Price, 2005).

5.2.2. Information obtained from CD studies of proteins

1) Secondary structure composition (% α-helix, β-sheets, turns): The absorption in this case is in the region of far UV of below 240 nm, mainly due to the peptide bond. The appearance of CD spectra in the far UV is contributed by the different types of regular secondary structures found in proteins. Secondary structure composition of a protein can be estimated by a number of algorithms that uses the data from far UV CD spectra. The algorithms used are SELCON, VARSLC, CDSSTR, CONTIN. In most of the procedures, basic datasets comprises the CD spectra of proteins whose structure has been solved by X-ray crystallography. Recently an online server called the DICHROWEB has been developed to analyse various CD spectra.

2) Tertiary structure determination: Amino acids tend to have a characteristic wavelength profile. Phenylalanine shows a band between 255 and 270 nm, tyrosine has a peak between 275 and 288 nm, whereas tryptophan has a peak between 290 and 305 nm. The near UV CD spectrum of a protein depends on various parameters like the

number of each type of aromatic amino acids present, their mobility, the nature of their environment (H-bonding, polar groups and polarisability).

3) Information about the overall structural features of proteins: The far UV CD analysis gives quantitative estimation of secondary structure and can be compared with those from X-ray crystallography and NMR. The structural relationships between native and recombinant proteins, wild type and mutant proteins can be assessed from the different CD spectral regions. Also the integrity of the expressed domains of a multi-domain protein can be confirmed from a CD data. The quantitative estimation of the stability of the folded state of a protein can be assessed, by the loss of CD signals either on increasing the temperature or by addition of denaturing agents.

4) Confirmational changes in proteins: Ligand binding causes structural changes in a protein, and is an important part in the mechanism of action and regulation of biological activity. CD spectroscopic analysis provides a mean of estimating such changes from its signals in different spectral regions. The range of ligand concentrations over which the structural changes take place, the extent of changes in the protein of interest, and the speed at which such changes occur all can be assessed from CD. When X-ray crystallographic analysis is not possible, CD can be used as an invaluable method in the study of protein structure.

CD is a demanding technique in terms of quantity of sample required for analysis or time of analysis. Good quality CD spectra can be obtained from less than 0.1 mg of protein sample (for far UV) and 1 mg (for the near UV). The technique is non destructive and time of analysis is 30 minutes or less (Kelly, Jess and Price, 2005).

An online server facilitating the analysis of circular dichroism spectroscopic data is the DICHROWEB. This is a very popular server and uses five important databases, and accepts the input data in a wide range of formats. It calculates the secondary structure contents and graphical analyses, comparing calculated structures and experimental data (Whitmore and Wallace, 2004; Lobley, Whitmore and Wallace, 2002).

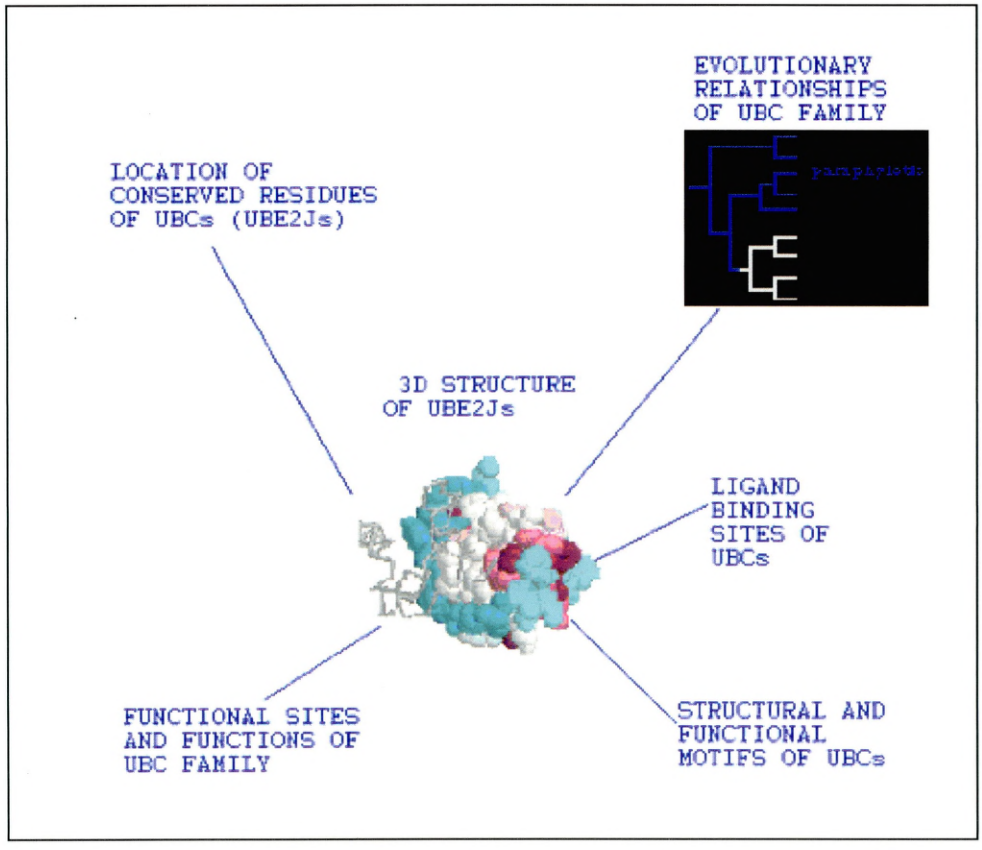
CHAPTER SIX

INTRODUCTION TO COMPUTATIONAL STRUCTURE PREDICTION (HOMOLOGY MODELLING)

6.1 The importance of 3D structures

The desire to understand biological processes at the molecular level, had led to the prediction of three dimensional protein structures from amino acid sequences, and hence the importance of protein homology modelling in addition to those of X-ray crystallography and NMR (Adams, Grosse-Kunstleve and Brunger, 2003). The structure prediction of a protein is essential for structure-based drug design, analysis of protein function and protein interactions. A fundamental principle in protein science is that "*protein structure leads to protein function*" (Scheef and Fink, 2003; http://www.biophysics.org/education/ellis.pdf).

Figure 6.1 Information that can be obtained from 3D structure of UBE2J



(This figure is adapted from Bartlett, Todd and Thornton, 2003)

Figure 6.1 summarises the biological information that can be obtained from an actual or predicted three dimensional protein structure.

The best method to find out the structure of a protein is either by NMR or by X-ray crystallography, but these methods have their limitations. Some proteins are either too large for carrying out its NMR analysis, or the protein is difficult to crystallize. Both of these methods are very demanding, and the involvement of more than one expertise, and it may not always be able to carry out the experiments economically. In these circumstances, the best method of structure prediction is computational homology modelling. The number of proteins in the PDB database is increasing everyday, by the constant input of NMR and X-ray predicted structures. Therefore to find homology of the sequence of protein whose structure needs to be predicted, to the desired percentage identity is not a problem. Theoretically the percentage identity of the two homologues (the template and the protein whose structure needs to be predicted) should be more than 30% (Rodriguez et al., 1998; Hung and Samudrala, 2003).

Figure 6.2 Homology zones from percentage identity of residues

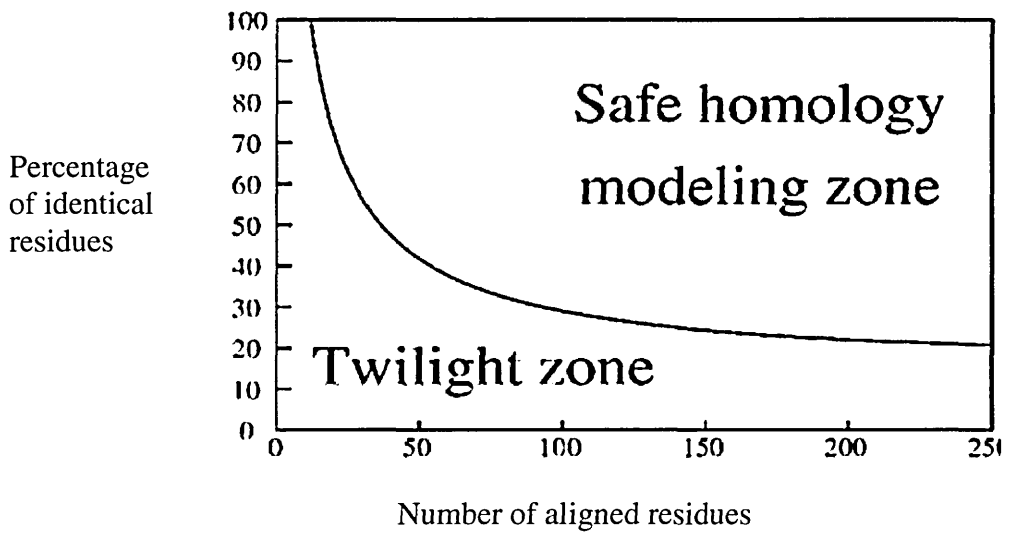


Figure 6.2 shows the safe homology modelling zone and the twilight zone. Two sequences are practically guaranteed to adopt a similar structure if the length of the two sequences and the percentage identity fall in the region marked as safe. This relationship was first identified by Chotia and Lesk (1986) (Figure adapted from Krieger, Nabuurs and Vriend, 2003).

Therefore, there is a high probability of trading a relative stringent watch between a known structure (template sequence) and that of a subject sequence. The predicted structure generated by homology modelling has to be treated cautiously but its utility can be demonstrated by the success of its use in other structure-dependent functions such as, docking of accessory molecules. In other ways, the predicted structure can serve as a useful starting point for trial-and-error investigation.

6.2. Relationship between structure and function

The relationship between structure and function can be examined on different levels like class, fold, homology, and analogy.

Protein structural class and enzyme function:

Structural classes of proteins form enzymes, of which α/β folds are over expressed in enzymes compared with a normal distribution. This over expression is due to the large number of nucleotide binding domain in enzymes. Associated with non-enzymes are the all- α and small folds. The transferases and hydrolases are particularly common among the α/β folds which reflect evolutionary selection.

Protein folds and function:

Proteins with similar structure can have totally different functions. This is because the number of folds is limited, and there are multiple super-families within each fold group. The most commonly found fold is the $(\beta/\alpha)_8$ barrel, which has reoccurred a number of times in evolution and has many diverse functions. The five most versatile folds in all proteins are the TIM barrel fold, Alpha/beta hydrolase fold, NAD binding domain, P-loop NTP hydrolase fold, Ferrodoxin like fold. They all are α/β or α + β folds, and each have different functions associated with it.

Homologous families and functions:

Within the homologous protein families it is expected that family members will have related functions, but this is not always the case. Considerable diversity has been seen within homologous super-families. Two proteins having high sequence identity between them could vary in their function. On the other hand during evolution, a protein family

might have been subjected to multiple amino acid changes, but their function has remained unchanged.

Analogues:

There are proteins that are analogous to one another sharing structural similarity but no sequence identity. Such analogues are examples of convergent evolution (Bartlett, Todd and Thornton, 2003).

CHAPTERS 7 – 10

METHODOLOGIES USED IN THIS RESEARCH PROJECT

CHAPTER SEVEN

METHODOLOGY (PHYLOGENETIC ANALYSIS)

7.1. Methods of the phylogenetic analysis of the UBC and UBC like peptide sequences.

Sequence data mining

Retrieving all known 13 yeast UBCs (UBC1-UBC13) peptide sequences from the EXPASY site (www.expasy.org). BLAST (Basic Local Alignment Search Tool) search using each of the 13 yeast UBC peptide sequences to find its homologues in different organisms whose genome was fully sequenced in appropriate databases NCBI, ENSEMBL, EBI, TIGR (www.ncbi.nlm.nih.gov, www.ensembl.org, www.ebi.ac.uk, www.tigr.org). The different organisms whose homologues were retrieved were *Homo sapiens, Mus musculus, Arabidopsis thaliana, Oryza sativa, Drosophila melanogaster, Caenorhabditis elegans, Plasmodium falciparum, Anopheles gambaie, Neurospora crassa, Schizosaccharomyces pombe and Plasmodium yoelli yoelli.*

All homologues retrieved by the BLAST search were given arbitrary names, as they were only indicated by accession numbers without any proper nomenclature naming. There were often repeat and partial sequences that had appeared in the database BLAST searches which were excluded.

More than one homologue in one organism was often retrieved from the blast search, each differing from the other by a different accession number. An example to cite this illustration is using yeast UBC13 peptide sequence to a BLAST search to find its homologues. Three homologues were obtained from the BLAST search for *Arabidopsis thaliana.* These homologues came up with different accession numbers, with no proper nomenclature naming. Arbitrary names were therefore given as AtUBC13a, AtUBC13b, and AtUBC13c. Phylogenetic analysis of those arbitrarily named peptide sequences were then carried out. It was from the branches of the phylogenetic tree, that the arbitrary namings were reconfirmed. For example, if the arbitrarily named peptide sequences appeared in the same phylogenetic branch of the same yeast UBC (e.g. AtUBC13a appearing in the same branch as ScUBC13), then it was confirmed to be a probable true homologue/orthologue of ScUBC13. But if AtUBC13a appeared in a different branch along with ScUBC10, then the AtUBC13a was renamed as AtUBC10a, since it is in the same branch as ScUBC10. A pairwise alignment was also carried out of AtUBC13a, first with ScUBC10 and then with ScUBC13 to reconfirm this, where the percentage identity was compared. The chromosomal location search was carried out of peptide sequences, to find out gene duplication or splice variants. From the

chromosomal location search, two or more duplicate genes were confirmed to be either a gene duplication or a splice variant of a single gene.

Multiple sequence alignment of the retrieved sequences

A multiple sequence alignment (MSA) of all retrieved peptide sequences were carried out. Initial alignment was carried out by the MSA tool CLUSTALW (www.ebi.ac.uk/clustalW). As the alignment generated by CLUSTALW was biologically not significant, it was further manually aligned by the alignment software tool GENEDOC to obtain the desired alignment.

The human TSG101, human CROC-1 (UBE2V1) and human UBE2V2 are non-UBC genes but they are UBC like. They lack the active site residue cysteine and the UBC PROSITE signature. So they were aligned as according to the alignments published by (Sancho et al., 1998; Ponting, Cai and Bork, 1997). The conserved region of the multiple sequence alignment was then selected and transferred to the UNIX platform for its phylogenetic analysis.

Phylogenetic analysis in Phylip 3.6

The phylogenetic analysis was carried out in the unix version of Phylip 3.6. There are two types of phylogenetic trees, rooted and unrooted. A rooted tree here gave the ancestral information (root) of the peptide sequences stating clearly which peptide sequence evolved first, the hierarchy of evolution and also the relationship of one sequence with the other. The steps involved in the phylogenetic analysis were as follows: Tree-puzzle (to calculate the Coefficient of Variation), Seqboot (to assign a statistical significance to the tree), Protdist and Neighbor-joining (the distance based method of phylogeny), consense (to derive a consensus of all the different trees that had been generated and reproduce 1 consensus tree out of it), Drawgram (to plot the rooted tree generated from consense) and Adobe Illustrator (plots the rooted phylogenetic tree from drawgram and all editing carried out).

Different steps in Phylip:

Tree-puzzle:

The "Tree-puzzle" programme was used to calculate the coefficient of variation (CV) from the invariant amino acids in the multiple sequence alignment.

$$CV = 1/\sqrt{\alpha}$$

α is the gamma distribution parameter which was obtained from the tree-puzzle analysis.

Phylip-seqboot

The "Seqboot" analysis was carried out by using the MSA in order to calculate bootstrap values for the phylogenetic tree. The number of replicates used were 100. The "outfile" generated was the result of the bootstrap analysis, which was used for the next step of analysis called "Protdist".

Phylip-protdist

"Outfile" of Seqboot was used for the "Protdist" analysis and 100 datasets were asked to be analysed here, also the CV value (that was calculated by the tree-puzzle analysis) was entered here in this analysis. The result obtained was the "outfile" which was used to carry out the next step of analysis called "Neighbor" (neighbour-joining method).

Phylip-neighbor

"Outfile" of Protdist was used for the "Neighbor-joining" analysis, where 100 datasets were also asked to be analysed. The results obtained from it was as "outfile and outtree", where the "outtree" was used to perform the consense (consensus of the 100 trees generate in the neighbour-joining step).

Phylip-consense

"Outtree" of Neighbor-joining analysis was used for the "consense" analysis. The result obtained from it was as "outfile and outtree", where the "outfile" consisted of the unrooted phylogenetic tree with the bootstrap values and the "outtree" was the other result which was used to draw the rooted phylogenetic tree.

Drawgram:

"Drawgram" was then used to plot the rooted phylogenetic tree from the previous phylip-consense analysis result "outtree". The edited phylogenetic tree was then saved as a postscript file. This post script file was then used to plot the rooted phylogenetic tree in any of the photo editing softwares.

Editing of the generated phylogenetic tree

Adobe-Illustrator:

The post-script file that was generated in the previous Drawgram analysis was finally used in the Adobe-Illustrator programme to plot the rooted tree for further diagrammatic representations and incorporating the bootstrap values at each node of the phylogenetic tree, which was finally saved as a PDF file with the help of the adobe Illustrator program.

Annotation of the phylogenetic tree

The arbitrary names that were given initially were then confirmed from its appearance in the respective branches of the phylogenetic tree. The displayed phylogenetic tree branches in the "Adobe Illustrator" were then annotated into functional and evolutionary groups, and also the nomenclature naming of all yeast and human UBCs were reviewed.

7.2. The phylogenetic analysis of drosophila $TAF_{II}250$, selected UBCs and human CROC-1 (UBE2V1)

Sequence data mining

The *Drosophila* $TAF_{II}250$ sequence was retrieved from the NCBI protein database (http://www.ncbi.nlm.nih.gov/). This sequence was used to carry out both a BLASTP and a TBLASTN search (http://www.ncbi.nlm.nih.gov/BLAST/) to identify its homologues in different organisms including some of the insects, vertebrates and nematodes. Using the known UBC active site, PROSITE signature pattern (http://www.expasy.org/prosite/), a manual search was carried out in the *Drosophila* TAFII250 sequence for all cysteine residues simply by using the find function in *notepad.* The UBC active site was then located in all homologues of *Drosophila* $TAF_{II}250$, but was only found to be present in *Drosophila pseudoobscura* and *Apis mellifera.*

Multiple sequence alignment

Following the identification of UBC active site in *Drosophila* $TAF_{II}250$ and few of its homologues, a multiple sequence alignment (MSA) was carried out in CLUSTALW (http://www.ebi.ac.uk/clustalw/index.html) using both complete and partial $TAF_{II}250$ sequences from different other species and also including human CROC-1. The *Drosophila* $TAF_{II}250$, together with all its homologous peptide sequences, also including the human CROC1 (UBE2V1) peptide sequence and some of the UBC(E2s) whose 3D structure is known, were then taken together in the "FASTA" format (the format which is required for MSA) and a multiple sequence alignment (MSA) was carried out in ClustalW. The putative $TAF_{II}250$ UBC active site region of the ClustalW generated MSA, was then manually aligned to the UBC active site using the manual alignment software GENEDOC (http://www.psc.edu/biomed/genedoc/). In addition CROC-1 was also manually aligned using previous UBC, CROC-1 and $TAF_{II}250$ alignments published by Ponting, Cai and Bork (1997) and Rothofsky and Lin (1997).

Also, the amino acid region 191- 214 of CROC1 and the amino acid region 1225- 1248 of $TAF_{II}250$ has 41.7% identity as indicated in the paper by Rothofsky and Lin (1997).

Hence these conserved regions were also manually aligned in the GENEDOC alignment program. The conserved regions were then used to carry out its phylogenetic analysis.

Phylogenetic analysis

The manual alignment generated in GENEDOC, was then saved in Phylip format for Neighbor-joining phylogenetic analysis using Phylip 3.6, as detailed in previous section 7.1 (http://evolution.genetics.washington.edu/phylip.html).

Phylip 3.6 program was used to carry out the phylogenetic analysis, using the Neighbor-joining algorithm, and DRAWTREE was used to plot the unrooted phylogenetic tree. The Adobe Illustrator was then used to complete all the editing to reproduce a well presentable phylogenetic tree with its bootstrap values. Drawtree and not Drawgram was used to plot the unrooted tree, because here the objective was to see the relationships of the peptide sequences, and not the root or the ancestral inference of the peptide sequences.

Phylogenetic tree annotation

The evolutionary and functional relationship of all $TAF_{II}250$s, UBCs and UBC like human CROC-1 (UBE2V1) were inferred from the generated phylogenetic tree.

CHAPTER EIGHT

METHODOLOGY ON CONSURF ANALYSIS

8.1 Methodology of consurf analysis

In the analysis of Consurf (http://conseq.bioinfo.tau.ac.il/) PDB ID 2F4WB (which is the structure of UBE2J2), was taken as the input for the query protein.

The next "optional" criterion of the analysis procedure was, to either let the Consurf find its homologous peptide sequences, or the multiple sequence alignment of its homologues could be submitted beforehand. Here the prealigned 193 peptide sequences (that was used for the phylogenetic analysi in phylip) was submitted to the database for analysis. A phylogenetic tree was constructed by Consurf, of the MSA using the neighbor-joining algorithm. The conservation of the query peptide sequence was calculated and these conservation scores that were plotted on the 3D structure of the query protein. The conservations were represented in different colours, divided into nine grades, where Grade-1 is the most variable position and is assigned the colour Turquoise. Grade-5 is the intermediate conservation, with a colour assigned white, and Grade-9 which is the most conserved is assigned the colour Maroon (Landau, 2005; Armon, Graur and Ben-Tal, 2001).

CHAPTER NINE

METHODOLOGY OF STRUCTURE PREDICTION BY X-RAY CRYSTALLOGRAPHY

A summary of the various steps that were involved in the production of protein for X-ray crystallographic determination of structure were as follows:

Cloning

Design of primers to amplify the UBE2J1 fragment,

Preparation of plasmids and PCR amplification of UBE2J1,

Digestion and ligation of PCR amplified fragment,

Transformation, purification of plasmid of *E. coli* BL21,

Sequencing of the cloned sample to confirm identity of sample,

Protein production and purification

Protein induction of transformed cells,

Cell rupture to release recombinant fusion protein,

Nickel binding of fusion protein to Ni-NTA column

Protein elution from Ni-NTA column,

Purification of the eluted and concentrated protein by size exclusion chromatography,

Protein analysis

Mass spectrometric analysis of the purified protein,

CD spectrum of the purified protein,

Crystallization

Crystallization of the purified protein,

X-ray diffraction

X-ray diffraction of the crystal obtained.

A human homologue of yeast UBC6, called UBE2J1 was used for crystallization. The whole of the protein was not used as there is a transmembrane domain region, which would be very difficult to crystallize (Tusnady, Dosztanyi and Simon, 2005). Multiple sequence alignment of UBE2J1 and all its homologues were carried out using ClustalW. The most conserved region around the active site was selected to crystallize. The region of UBE2J1 that was used is as shown in Figure 9.1. The peptide fragment contains 170 amino acids and has a molecular weight of 19 KDa.

Figure 9.1. Portion of the UBE2J1 peptide sequence used to design the primers

METRYNL**KSPAVKRLMKEAA ELKDPTDHYH AQPLEDNLFE WHFTVRGPPDSDFDGGVYHGRIVLPPEYPMKPPSIILLTANGRFEVGKKI CLSISGHHPETWQPSWSIRTALLAIIGFMPTKGEGAIGSLDYTPEERRAL AKKSQDFCCEGCGSAMKDVLLPLKSGSDSSQADQEAKEL**ARQISFKAEV NSSGKTISESDLNHSFSLTDLQDDIPTTFQGATASTSYGLQNSSAASFHQPTQP VAKNTSMSPRQRRAQQQSQRRLSTSPDVIQGHQPRDNHTDHGGSAVLIVILT LALAALIFRRIYLAN EYIFDFEL

Figure 9.1 is the whole UBE2J1 peptide sequence, with the transmembrane domain SAVLIVILTLALAALIF and the UBC6 PROSITE domain ***T-[PAR]-[NS]-G-R-F-x(3)-[KTE]-[RK]-[LIV]-C-[LMS]-[ST]-[IMF]-[ST]-x(2)-H-[PK]*** *underlined in bold. The peptide fragment that had been used to crystallise, is in bold. UBE2J1 was cloned in between the 5' BamH1 and 3' EcoR1 restriction sites of pHISTEV30a plasmid vector.*

Table 9.1 The whole of the nucleotide sequence of UBE2J1 that was used to design the primers.

1	AAGAGTCCGG CTGTTAAACG TTTAATGAAA GAAGCGGCAG AATTGAAAGA
51	TCCAACAGAT CATTACCATG CGCAGCCTTT AGAGGATAAC CTTTTTGAAT
101	GGCACTTCAC GGTTAGAGGG CCCCCAGACT CCGATTTTGA TGGAGGAGTT
151	TATCACGGGC GGATAGTACT GCCACCAGAG TATCCCATGA AACCACCAAG
201	CATTATTCTC CTAACGGCTA ATGGTCGATT TGAAGTGGGC AAGAAAATCT
251	GTTTGAGCAT CTCAGGCCAT CATCCTGAAA CTTGGCAGCC TTCGTGGAGT
301	ATAAGGACAG CATTATTAGC CATCATTGGG TTTATGCCAA CAAAAGGAGA
351	GGGAGCCATA GGTTCTCTAG ATTACACTCC TGAGGAAAGA AGAGCACTTG
401	CCAAAAAATC ACAAGATTTC TGTTGTGAAG GATGTGGCTC TGCCATGAAG
451	GATGTCCTGT TGCCTTTAAA ATCTGGAAGC GATTCAAGCC AAGCTGACCA
501	AGAAGCCAAA GAACTG

Primers were used to incorporate *BamH1* at the 5' end and EcoR1 at the 3' end. The forward primer was designed to incorporate a *BamH1* restriction enzyme as follows:
3 extra nucleotides (CGC) were added to generate a BamH1 site.

Note that the 3 extra nucleotides were added to the 21 nucleotides of UBE2J1 to incorporate a *BamH1* site.

5' CGCGGATCCAAGAGTCCGGCTGTTAAACGT 3'

In the reverse primer also 3 extra nucleotides were added to the 21 nucleotides of UBE2J1 to incorporate an *EcoR1* site.

GACCAAGAAGCCAAAGAACTG ---- is the 21 nucleotide sequence in the 5' to 3' with the addition of 3 stop codons (TAA TAG TGA) and 3 extra nucleotides (CCG) to generate an *EcoR1* site.

5' CCGGAATTCTCACTATTACAGTTCTTTGGCTTCTTGGTC 3'

9.1. PCR reactions

9.1.1 The PCR mix

As the concentrations of the primers were 100 pmol/µl, they were diluted to 10 pmol/µl for the PCR reaction. The cDNA, purchased from BioChain Institute, USA, was isolated from the human brain tissue.

The PCR-mix includes 12.5 µl PCR Master mix 2X (50 units.ml TaqDNA polymerase, 400 µM each of dATP, dGTP, dCTP, dTTP and 3mM $MGCL_2$), 2.5 µl UBE2J1 (Forward), 2.5 µl UBE2J1 (Reverse), 1µl cDNA and 6.5 µl Nuclease free water.

A control PCR mix was also prepared without the cDNA.

9.1.2. Calculation of annealing temperatures

To calculate the annealing temperatures, the melting temperatures (t_m) of both the forward and the reverse primers were calculated separately. All G and C nucleotides were given 4°C and 2°C for every A and T. From the calculated t_m, 5° was subtracted, which was used as the annealing temperature for the first 10 cycles. In the next 20 cycles the temperature was increased so as to allow the further binding of the PCR products and not allow further binding of the PCR product to the cDNA.

9.1.3 Setting of the PCR machine

The lid of the PCR machine was preheated for 4 minutes at 105°C. This was followed by an initial denaturation set for 2 minutes at 94°C. The sample was then placed into the PCR machine with the first 10 cycles of the following: 94°C for 30 seconds, 50°C for 30 seconds and 72°C for 30 seconds. The next 20 cycles of the following were repeated as: 94°C for 30 seconds, 59°C for 30 seconds, 72°C for 30 seconds. The reaction was

completed by a final chain extension step for 5 minutes at 72°C followed by a holding step at 10°C.

After the completion of the PCR reaction the products were analysed by agarose gel electrophoresis.

9.2. DNA preparation

The DNA preparations (maxi-preps, mini-preps, and gel extractions) used for cloning were carried out with Qiagen™ kits, in accordance with the manufactures instructions. DNA restrictions enzymes were obtained from both Promega™ and New England Biolabs™ (NEB).

Preparation of LB media:

To make 1 litre of LB media, 1% w/v bacto-tryptone, 0.5% w/v bacto-yeast extract and 0.5% w/v NaCl were added to 800 ml water. The pH was adjusted to 7.5 with NaOH, then 1% w/v of Agarose was added to the mixture and the final volume made to 1 litre and then autoclaved to sterilise. The autoclaved medium was allowed to cool to 55° C and finally kanamycin 50 $\mu g\ ml^{-1}$ was added.

9.3. Transformation of plasmids

The plasmids were kindly supplied by Dr. Dimitris Xerodimus of Prof. Ron Hay's laboratory of the University of St. Andrews.

pHISTEV30a is detailed in Figure 9.2a and pHISTEV30a-GST-thrombin is detailed in Figure 9.2b.

Figure 9.2.1 pHISTEV30a plasmid vector

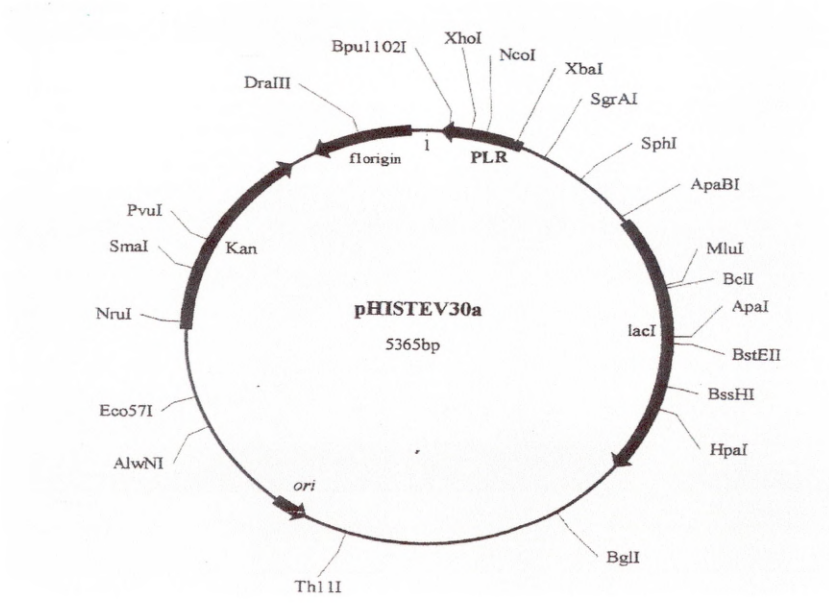
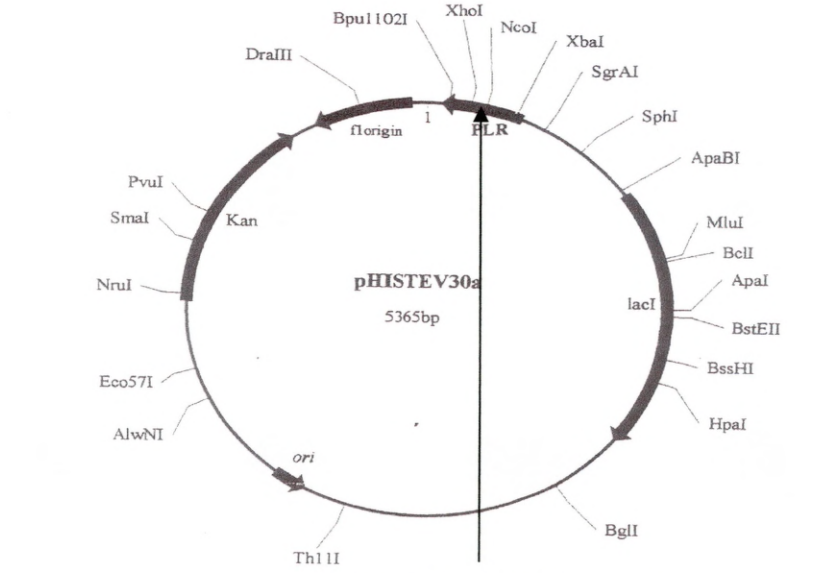


Figure 9.2.2 pHISTEV30a-GST-thrombin plasmid vector



pHISTEV30a-GST-thrombin

Figure 9.2.1 is the plasmid vector pHISTEV30a where the UBE2J1 fragment was inserted between the BamH1 and EcoR1 which lies within Nco1 and Xho1 of the plasmid vector.

Figure 9.2.2 the plasmid vector pHISTEV30a-GST-thrombin where GST-thrombin has been cloned in the Nco1 and Xho1 region of the plasmid vector.

Figure 9.3.1 Plasmid map showing the cloning site of the UBE2J1 fragment

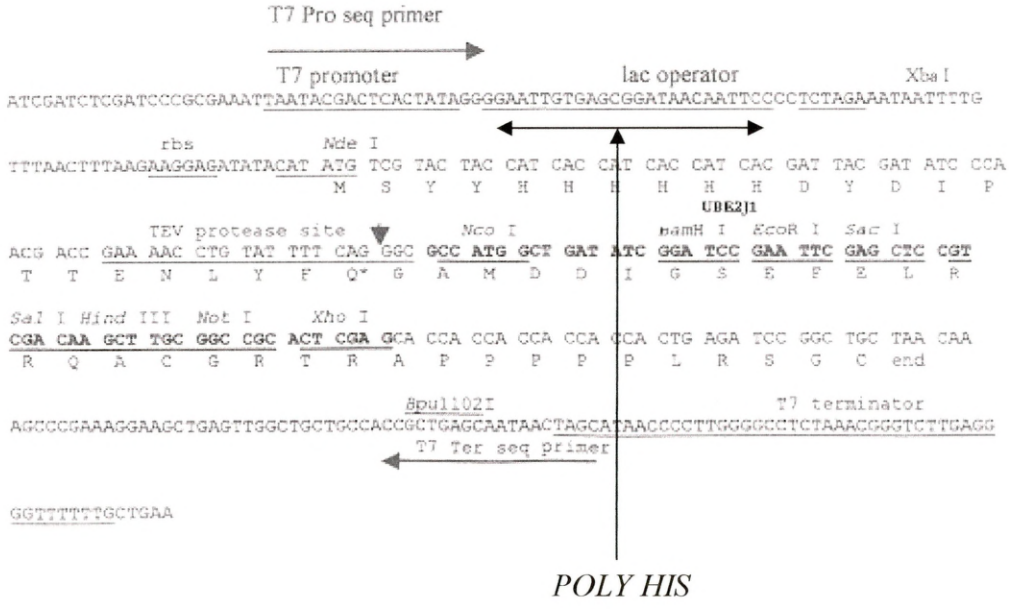


Figure 9.3.1 above is an expansion of the plasmid map of figure 9.2.1, which shows the site into which UBE2J1 fragment was cloned in between the BamH1 and EcoR1.

Figure 9.3.2 Plasmid map showing the cloning site of GST-thrombin

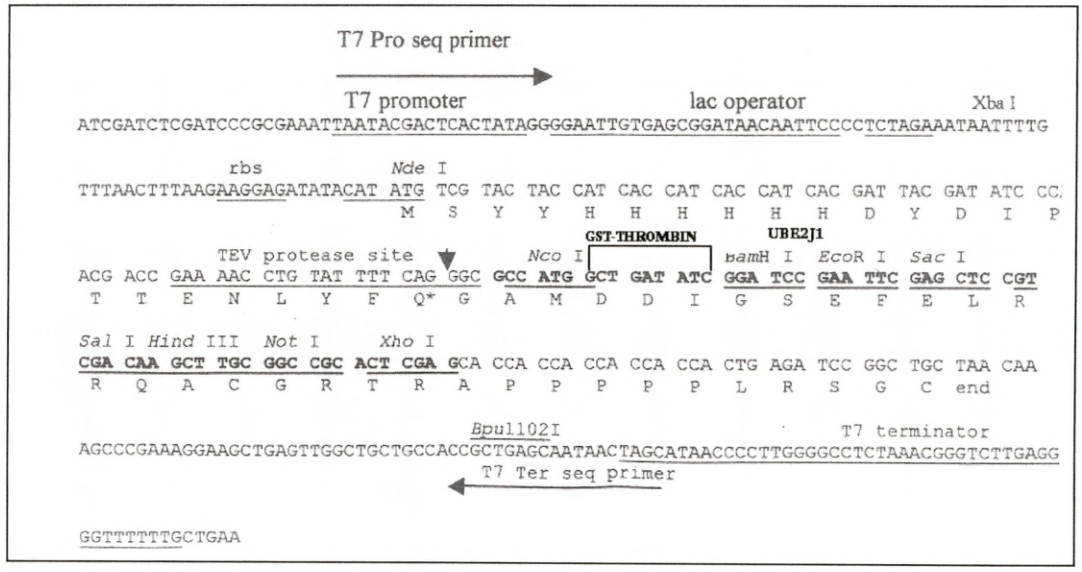


Figure 9.3.2 above is an expansion of the plasmid map of figure 9.2.2, which shows the site into which GST-Thrombin was cloned by Prof. Ron Hay's group in between the Nco1 and the BamH1, which is just before the UBE2J1 cloned site.

Procedure for the transformation of plasmids

Plasmid GST-HIS-TEV or HIS-TEV (1µl) and 50 µl of JM109 competent cells were mixed, then incubated on ice for 20 minutes. Heat shock was applied to the cells at 42°C for 45 – 50 seconds and the mixture was then again placed on ice for 2 minutes.
SOC medium 950 µl [2% w/v Bacto-tryptone, 0.5% w/v bacto-yeast extract, 0.01mM NaCl, 0.0025mM KCl, 0.005mM Mg^{2+} stock and 0.005mM glucose., adjusted to pH 7 and volume made up to 100 ml. After autoclaving, a final concentration of Mg^{2+} (0.005mM) and glucose (0.005mM) were added] was added with gentle mixing and incubated at 37°C for 1½ hours. An aliquot (100 µl) of this mixture was applied to a kanamycin plate and spread using the sterile glass rod and incubated at 37°C for 18 – 24 hours.

9.4. Plasmid preparation

One colony each of GST and HIS-TEV plasmids was grown at 37°C overnight in 5ml LB media with Kanamycin (50 µg ml^{-1}). 2ml of this overnight growth was inoculated into 500ml LB media with 500 µl Kanamycin(50 µg ml^{-1}) in a 1litre flask. This was incubated at 37°C overnight. The bacterial cells were then harvested by centrifugation for 5 minutes at 4000 x g and plasmids isolated using the "QIAGEN MAXI PREP." protocol and kit. Plasmid DNA was quantified spectrophotometrically.

QIAGEN MAXI PREP procedure:

The bacterial cells were harvested by centrifugation at 4000 x g at 4°C for 15 minutes and resuspended in 10 ml. buffer P1 (50 mM Tris HCl pH 8.0, 10 mM EDTA and 10 ug.ml RNASE A). Buffer P2 (10 ml) (200mM NaOH, 1% SDS, 1M NaOH 200 ml, 10% SDS 100 ml, water 700 ml) was added, mixed gently but thoroughly by inverting 4-6 times and incubated at room temperature for 5 minutes.
Chilled buffer P3 10 ml [3.0 M KOAc (potassium acetate) pH 5.5 (29.5% w/v in water)] was added, mixed immediately but gently by inverting 4-6 times and incubated on ice for 20 minutes. The mixture was centrifuged at 3000 x g for 15 minutes at 4°C and supernatant containing plasmid DNA was removed promptly,
A QIAGEN-tip 500 was equilibriated by applying 10 ml buffer QBT (4.38% w/v NaCl, 1.04% w/v MOPS dissolved in 800 ml distilled water, pH adjusted to 7 with NaOH, 15%v/v ml isopropanol and 15 ml 10% Triton X-100 solution added and volume

adjusted to 1 litre with distilled water) and the column was allowed to empty by gravity flow. A filter was put on top of the QIAGEN tip and the supernatant was applied to pass through the filter and allowed to enter the resin by gravity flow.

The QIAGEN tip was washed with 2 x 30 ml buffer QC (5.84% w/v NaCl, 1.04% w/v MOPS dissolved in 800 ml distilled water, pH adjusted to 7 with NaOH and then 15% v/v isopropanol added. Finally distilled water was added to adjust the volume to 1 litre). DNA was eluted with 15 ml buffer QF (7.30% w/v NaCl and 0.60% w/v Tris base, pH 8.5 with the addition of 15% v/v isopropanol) and was precipitated by adding 10.5 ml room temperature isopropanol to the elutant. The mixture was centrifuged at 4°C for 30 minutes at 10,000 x g. The removed pellets were washed with 5 ml of room temperature 70 % v/v ethanol and centrifuged at 10,000 x g for 10 minutes. The supernatant was decanted without disturbing the pellets, which were air dried for 5-10 minutes and the DNA redissolved in a suitable volume of distilled water.

9.5. Growing of the GST-HIS-TEV and HIS-TEV colonies from the kanamycin plates

Kanamycin 50 mg ml^{-1} (5 µl) was added to 10ml of LB media and inoculated with a single colony from the kanamycin plate (using a micro-pipette tip) then incubated at 37°C for 12 – 16 hours. Cell growth was checked visually, and the plasmids were purified using the QIAGEN maxi prep detailed in section 9.4 and stored in 70% glycerol at -70°C.

9.6. Gel electrophoresis

Agarose gel 1.2 % w/v was prepared with 1 X TBE and 10 µl of 0.0005% w/v ethidium bromide. DNA marker (10 µl of a 100bp DNA ladder) was loaded onto the first lane and 60 µl (50 µl of each of the two PCR product and control + 10 µl of loading dye) of the experimental mix was loaded onto the other lanes. 1 X TBE (Tris-Borate-EDTA) buffer was used for running the gel.

9.7. Digestion of the PCR product of UBE2J1 fragment and the plasmids

PCR products and the plasmids were both digested with BamH1 and EcoR1 using the following incubation mixture: 5µl of 10 X Buffer H (composition of 900mM TrisHCL pH 7.5, 500mM NaCl, and 100 mM $MgCl_2$ at 37°C), 0.5 µl BSA (0.5% w/v), 1.5 µl

EcoR1 (10mM Tris-HCl (pH 7.4), 400mM NaCl, 0.1mM EDTA, 1mM DTT, 0.15% TritonX-100, 0.5mg/ml BSA, 50% glycerol), 1.5µl BamH1 (10mM Tris-HCl (pH 7.4), 300mM KCl, 5mM $MgCl_2$, 0.1mM EDTA, 1mM DTT, 0.5% w/v BSA, 50% v/v glycerol), 10 µl DNA (PCR product), 31.5 µl nuclease free water. This reaction mixture was incubated at 37°C for 4 hours for complete digestion.

9.8. Gel extraction

The digested PCR product and the plasmids were cut from the gel bands and the DNA extracted using the QIAGEN kit. After cutting the band from the gel containing the DNA the gel slice was weighed. To 1 volume (weight) of the gel slice 3 volumes of QG buffer (http://www1.qiagen.com/literature/handbooks/PDF/DNACleanupAndConcentration/QQ_Spin/1021422_HBQQSpin_072002WW.pdf) was added and resuspended by vortexing for 30 seconds. It was then incubate for 10 minutes at 50°C and 1 gel volume of isopropanol was added and mixed. The sample was then centrifuged for 1 minute, the supernatant was removed and 0.5 ml of buffer QG was added to the spin column and centrifuged for 1 minute. The cells in the column was then washed by adding 0.75 ml of buffer PE (http://www1.qiagen.com/literature/handbooks/PDF/DNACleanupAndConcentration/QQ_Spin/1021422_HBQQSpin_072002WW.pdf), centrifuged for 1 minute. The supernatant was again removed and was centrifuging for an additional 1 minute at 17,900 x g. Finally the DNA was eluted by adding 50 µl of water and centrifuging for 1 minute. As the buffers QG and PE are company patents, its ingredients were not known.

9.9. PCR product and plasmid ligation reaction

The digested PCR product and the plasmids were ligated, the reaction mixture is as follows: 13µl PCR product, 13µl plasmid Vector, 3µl 2X Rapid ligation Buffer (60mM Tris-HCl (pH 7.8), 20mM MgCl2, 20mM DTT, 2mM ATP, 10% polyethylene glycol), 1µl T4 DNA Ligase (concentration is 3 weiss units/µl in 50mM Tris-Hcl, 10mM $MgCl_2$, 10mM DTT, 1mM ATP & 25 µg/ml BSA at pH 7.5 and 25°C). This ligation reaction was incubated overnight at 4°C for complete ligation.

9.10. Transformation of the ligated product in DH5α cells

The transformation reaction was carried out on ice for 30 minutes in 15 ml falcon tubes using the following incubation mix: 20µl Ligated product, 20µl 5X KCM [1x KCM is (100 mM KCl, 30mM of $CaCl_2$, 50 mM $MgCl_2$)], 60µl Water and 100µl of DH5α cells. After incubation LB media (1ml) was added to the mix and incubated at 37°C on an orbital shaker for 1 hour. Finally 100µl of this mix was spread on kanamycin plates that were incubated at 37°C overnight.

Single colonies were selected from the kanamycin plate and incubated in 10 ml LB kanamycin (50 µg ml^{-1}) overnight at 37°C with agitation. Cells were harvested by centrifugation for 15 minutes at 6000 x g and plasmid DNA extracted using the QIAGEN MAXI PREP as in section 9.4.

9.11. Digestion of ligated product

Both the PCR products and the plasmids were digested with the restriction enzymes BamH1 and EcoR1 which is as follows: 5µl 10 X Buffer H, 0.5 µl BSA, 1.5 µl EcoR1, 1.5µl BamH1, 10 µl DNA & 31.5 µl NF water. This was incubated at 37°C for 4 hours for complete digestion.

9.12. Preparation of BL21 competent cells

BL21 competent cells were required for the transformation of the ligated products. The BL21 competant cell were streaked out from the glycerol stock into agar plates to grow at 37°C overnight. A single colony was used to innoculate 10 ml LB media and this was incubated at 37°C overnight. The overnight culture (5ml) was used to inoculate 500 ml LB in a 2 litre baffled flask and was incubated at 37°C until OD_{600} approx. 0.6. The cells were centrifuged (2000 x g) for 5min at 4°C, were gently resuspend in 25ml cold TSB (LB pH 6.1 with HCl, 5% DMSO v/v, 10 mM $MgCl_2$ & 10 mM $MgSO_4$), incubated on ice for 10 min and was aliquoted into several eppendorfs, stored at -70°C.

9.13. Transformation of the ligated product into the BL21 competent cells

The ligated sample was transformed into the BL21 strain by mixing 9µl 1X KCM, 1µl DNA & 10µl BL21 cells. A control transformation was carried out to test the BL21 competant cells for any contamination. The test and the control were incubated on ice for 10 minutes and then at room temperature for 10 minutes, and finally plated out into

kanamycin plates, then incubated at 37°C overnight. A colony of each growth plate was inoculated into 10ml of LB media having kanamycin and allowed to grow at 37°C overnight.

9.14. DNA sequencing to confirm the constructed primers

The designed primers and the double stranded plasmids were sent for sequencing to the University of Dundee. The nucleotide sequences that were obtained from the chromatogram sent by the sequencing department were transcribed to the peptide sequence which was found correct (data not shown).

9.15. Protein induction

Overnight growth culture (10ml) of the BL21 transformed ligation product was inoculated each into a 1 litre LB media which had the required antibiotic kanamycin. A total amount of 3 litres was allowed to grow at 37°C, until an Optical Density (OD at 595 nm) of 0.6 – 0.8 was obtained. The solution was then cooled on ice and 0.25mM IPTG was added into the protein sample for induction at room temperature with constant orbital shaking for 5 to 6 hours. The culture was then centrifuged at 15,000 x g for 10 minutes to collect the cells separated from the supernatant, resuspended in 10ml of PBSA buffer, further centrifuged at 15,000 x g for 10 minutes and cells stored at -20°C.

9.16. Cell lysis

The induced cells were resuspended in 100ml of lysis buffer (50mM Tris pH 7.5, 0.5 M NaCl, 1% Triton 10 % & 1 Protease Inhibitor cocktail tablet in PBSA). The cell membranes were ruptured and the protein released (http://www.promega.com/pnotes/86/11217_23/11217_23.pdf).

The re-suspended cells in the lysis buffer were sonicated 4 times 15 seconds each at an amplitude of 15 and were then centrifuged for 15 minutes at 40,000 x g twice, so as to get a clear supernatant, which was then collected and a final volume concentration of 10mM β-mercaptoethanol was added.

9.17. SDS-PAGE analysis

Protein samples were resuspended in disruption buffer (1X: 20mM Tris/Hcl pH 6.8, 2% w/v SDS, 5% w/v β- mercaptoethanol, glycerol and 2.5% w/v bromophenol blue) and denatured at 100°C for 5 minutes, before loading on the SDS polyacrylamide gel (acrylamide percentage appropriate for the size of proteins to be separated). New England Biolabs™ prestained molecular weight markers were used as standards to establish the apparent molecular weights of proteins resolved on SDS- polyacrylamide gels. Separated polypeptides were stained with Coomassie Brilliant Blue (0.2 % w/v Coomassie Brilliant Blue R250; 50% v/v methanol; 10% v/v acetic acid) for thirty minutes and then destained (20% v/v methanol; 10% v/v acetic acid) overnight.

Procedure for gel preparation for SDS-PAGE

Running gel (12 % w/v) was prepared by adding 3.3ml H_2O, 4.0ml 30 % w/v Acrylamide mix, 2.5ml 1.5 M Tris (pH 8.8), 0.1ml 10% w/v SDS, 0.1ml 10% w/v APS & 0.004ml TEMED. Stacking gel was prepared by adding 6.8ml H_2O, 1.7ml 30 % w/v Acrylamide mix, 1.25ml 1.0 M Tris (pH 6.8), 0.1ml 10% w/v SDS, 0.1ml 10% w/v APS & 0.02ml TEMED.

9.18.1. Preparation of nickel beads

Ni-NTA Magnetic Agarose Beads were stored as a 5% w/v suspension in 30% v/v ethanol at 2-8°C. Prior to use beads were washed with 10 column volumes of (50 mM Tris (pH 7.5), 0.25M NaCl & 30 mM Imidazole in PBSA buffer).

9.18.2. Recharging of the nickel- NTA beads

The column was washed 3 times the column volume with water, then it was washed 3 times the column volume with 6M GuHCl and 0.2 M Acetic acid. It was then washed with 3 times the column volume with 100 mM $NiSO_4$. The nickel-NTA beads were again washed with water and then with 6M GuHCl and 0.2 M Acetic acid and finally stored at 4°C in 30% v/v ethanol.

9.19. Binding with the protein

The protein supernatant obtained in 9.16 was suspended with nickel beads and 10mM Imidazole and incubated with gentle agitation for 2 hours. The supernatant-nickel beads

mixture was then decanted into a column (approx 25 cm), and the supernatant eluted and the flow collected and stored at 4°C. The column was washed in 3 column volumes of wash buffer (50 mM Tris pH 7.5, 0.25M NaCl & 30 mM Imidazole in PBSA) and the flow-through retained and stored at 4°C. Elution buffer 8 ml (50mM Tris-pH 8.0, 500mM NaCl and 250mM Imidizole in PBSA) was applied to the nickel column and left for 10 minutes at room temperature. The first 5 ml fraction was collected and further elution buffer was run through and further 5 ml fractions were collected and assayed for the presence of protein (using a micro Bradford assay) until no more protein was eluted. The eluted fraction containing protein was dialyzed in 4 litre dialysis buffer overnight (50mM Tris pH 8.0 and 500mM NaCl in water) to remove Imidazole. The dialysed fraction was then centrifuged at 700 x g to remove any precipitate and DTT (1mM) was added.

9.20. Micro Bradford assay

Bradford reagent (100µl) was placed on each well of an Elisa plate. A drop was collected from the eluted fractions at regular intervals and mixed thoroughly with the Bradford reagent contained in the wells of the Elisa plate. A change in colour from reddish brown to blue indicated the presence of protein.

9.21. Quantitation of the eluted fraction containing protein

Protein concentration was determined using Bradford's method (Bradford, 1976). Protein samples were mixed with Bradford's (Biorad™) and the absorbance at 595 nm was measured on a spectrophotometer (Hitachi U-1100). Protein absorbance was calculated to mg/ml concentrations using a standard bovine serum albumin (BSA). Absorbance (at 595nm) of the mixture of 800 µl water, 20 µl BIORAD reagent and each of 2µg, 5µg and 10µg final concentration quantities of the standard (BSA) were taken and 2µl, 5µl and 10µl of the eluted fraction were taken and analysed, to estimate the amount of protein.

9.22. TEV protease clevage

After the quantification of the eluted protein as in section 9.19, typically a range of TEV protease concentration was incubated with 2µl protein (6.6 mg/ml) and the level of digestion determined using SDS-PAGE. The most effective concentration for complete

HIS-tag cleavage was taken to carry out TEV-protease cleavage of the whole batch of the eluted protein. The required amount of TEV protease was incubated for 5 hours with the protein, which was further for 2 hours with nickel beads and finally passed through the column, where the His-tag was retained with the nickel and the His-tag cleaved protein was collected as a flow through.

9.23 Concentrating the cleaved protein before gel purification

The cleaved protein was concentrated by centrifuging at 2647 x g using a Centricon Plus-20 centrifugal membrane filter which had a molecular weight cut off limit of 10,000.

9.24. Purification of the cloned, expressed UBE2J1 fragment by gel permeation chromatography

The buffer used for gel filtration chromatography contains 50mM Tris (pH 8.0), 500 mM NaCl in final volume in water, to which (1mM) DTT final concentration was added.

Column used for the gel permeation chromatography was Superdex 75, of 60 cm in length and 1.6 cm in diameter, Flow rate of the buffer 1.5 (ml/min), Peak fraction volume is 2.0 ml. Before the purification step, the column was equilibrated with the buffer.

9.25. Concentrating and desalting the protein

The gel purified protein was then concentrated to a final volume of 700µl by centrifuging at 2647 x g using the Centricon Plus-20 membrane filter which had a molecular weight cut off limit of 10,000.

As 500mM NaCl (in final concentration) generally interferes in the crystallization process, its final concentration was brought down to 100 mM by adding 10 ml of the dialysis buffer used as in section 9.24 (with the exception of the salt concentration being 100mM in the final volume) to the concentrator, while concentrating the protein. Typically 5 dilutions and centrifugations were used.

9.26. Crystallization

Crystal trials were carried out in the Robot (Hampton research crystal robot). For crystal trial 17 plates were set, each plate consisting of 96 wells were set for the crystallization of the protein UBE2J1. Around 200 µl (15.7 mg/ml) of the purified protein sample was used, where the machine aliquoted 100 nano-litres of the sample into each well for crystal trial. The various precipitants that were used are given in the appendix.
After aliquoting the precipitants by the machine, the plates were set in the incubator, for the crystal growth. Each of the plates were photographed by the machine everyday to check for the crystal growth.

9.27. X-ray diffraction of the crystal obtained

Theoretically the crystal grown should be optimised in the same condition the crystal was obtained, to get a much better crystal in size and form. In this case as the crystal obtained was suspected to be a salt crystal, the optimization procedure was not carried out, but the crystal was set for its x-ray diffraction. The crystal was first separated by a loop, and then mixed with the cryo-protectant consisting of PEG (polyethelene glycol) solution and mounted onto the goniometer, where a continuous stream of liquid nitrogen was passed through the loop having the crystal, so as to freeze the crystal and prevent it from denaturation by the X-ray beam. Finally the diffraction pattern was obtained, which confirmed that it was actually the crystal of the salt, and upon checking the precipitant used for that particular condition, it was found to be an ammonium sulphate crystal.

9.28. Circular dichroism (CD) spectrum analysis

The circular dichroism spectrum analysis was carried out at the St. Andrews University, the spectroscopic data obtained from the CD spectral analysis was then fed into the online server called as DICHROWEB, for the analysis of its secondary structure.

PARAMETERS USED in CD spectrum analysis

0.5 mg/ml concentration of the protein sample was required for the Far UV using a cell of pathlength 0.02 cm. 1 mg/ml concentration was used for the near UV using a cell of 0.5 cm.

The wavelength range of the far UV used was 190 – 260 nm, and that used for near UV was 250 – 320.

Table 9.2

Experimental parameters of CD spectrum analysis

Parameters	Far UV	Near UV
Concentration of the protein sample	0.5 mg/ml	1mg/ml
Cell path length	0.02cm	0.5cm
Wavelength range	190 – 260 nm	250 – 320nm
Number of scans	3	3

CD supports various algorithms like SELCON3, CONTINLL, CDSSTR VARSLC, K2d. The CDSSTR algorithm was used presently for the analysis. The data-files of the CD spectrum required for the input into the DICHROWEB are Aviv, Jasco, free format, Brookhaven SRCD, DRS, wherein Jasco was used here in the analysis. The data inputs are either in delta epsilon (which was used here), mean residue ellipticity, theta (machine units), or SRCD counts. The various wavelength intervals of the input data are 0.1, 0.2, 0.5 and 1nm.

CHAPTER TEN

METHODOLOGY OF COMPUTATIONAL STRUCTURE PREDICTION (HOMOLOGY MODELLING)

10.1 Methodology of homology modelling of UBE2J1 by DeepView

The Swiss-model server was developed by (Guex and Peitsch, 1997), and is one of the most preferred online homology modelling server (http://www.expasy.org/swissmod/). The Swiss-PDB viewer enables the analysis of proteins like homology modelling, where proteins can be superimposed to generate structural alignment, which compares relevant parts like active sites, amino acid mutations, hydrogen bonds, bond angles and distance between atoms, which are displayed by graphics and menu interfaces.
Here structural templates could be either given by the user or Swiss-PBD has the inbuilt option where it finds out the best template of the protein whose model needs to be generated. After the identification of the most suitable template, the next step was to superimpose the template structure to the target sequence. Following the superimposition a multiple sequence alignment of the template with that of the target sequence was generated. The best aligned template and the query sequence were then submitted to the Swis-PDB server for generation of the 3D structure. After obtaining the model from the Swiss-model server, an energy minimization step was carried out. The quality of the model was verified by the WHATCHECK report that was provided by DeepView. The model quality check was also obtained from various other parameters like colouring the model by B- factor, secondary structure, alignment diversity, and solvent accessibility (BBSRC Bioscience IT Services. 2003; http://www.usm.maine.edu/~rhodes/SPVTut/index.html).

10.2 Methodology of homology modelling of UBE2J1 by 3D-JIGSAW

The model of UBE2J1 generated by the online homology modelling server called 3D-JIGSAW was obtained by submitting the peptide sequence of UBE2J1 online to the server (http://www.bmm.icnet.uk/servers/3djigsaw/) and the model was generated by 3D-JIGSAW. The generated model file was opened up in Deepview and seen in Deepview and its Ramachandran plot obtained. The Whatcheck report was also obtained by submitting the model to Whatif database (http://swift.cmbi.ru.nl/WIWWWI/).

CHAPTERS 11 – 15
RESULTS AND DISCUSSION

CHAPTER ELEVEN

DESIGNING OF THE PROSITE SIGNATURE OF UBC6 FAMILY FROM THE MULTIPLE SEQUENCE ALIGNMENT (MSA)

11.1 Results and discussion of MSA & PROSITE signature

Thirteen distinct UBCs have been identified in yeast and 24 in humans (excluding all possible splice variants and those lacking a cysteine at the active site). All known UBCs have a core domain of around 150 amino acids. This so-called UBC domain contains the highly conserved UBC motif, which contains a highly conserved cysteine residue involved in thioester formation. The primary structure around the active site is highly conserved and has the consensus [FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x-[LIV] in virtually all UBCs, and forms the basis for the PROSITE signature for this class of enzyme (Figure 11.1). All UBCs containing this PROSITE signature contain the active site residue as cysteine. Two notable UBC families exist that do not conform to this canonical PROSITE signature. The first example is a family of so-called non-catalytic UBCs, which significantly lack the highly conserved cysteine at the UBC active site to which the ubiquitin molecule is attached, they are TSG101 and CROC-1 (UBE2V1) (Koonin and Abagyan, 1997; Xiao et al, 1998).

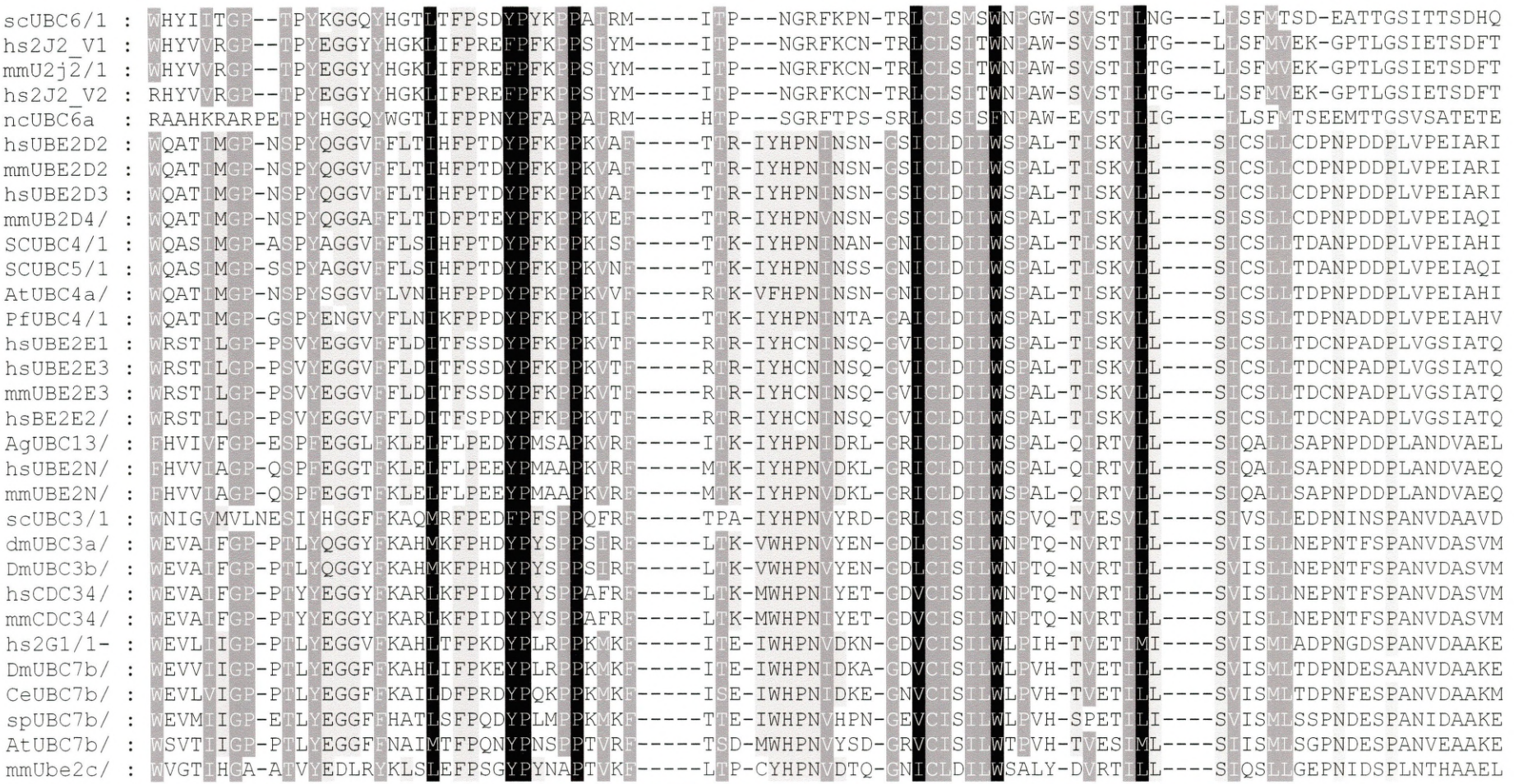
Secondly, there exists a family of yeast UBC6 enzymes and their homologues (Lester et al., 2000), which while having the highly conserved cysteine, found in most other UBC's, have a very different sequence from the existing UBC PROSITE signature. For example UBC6 lack the invariant histidine and glycine residues and has the extra amino acids G, R, F, which is not present in any other UBCs (Figure 11.1).

The PROSITE signature of all UBCs except UBC6 family is available in the database, but to date no PROSITE signature of the UBC6 family is available in the database (www.expasy.org). Therefore in order to create a PROSITE signature for the UBC6 family, an MSA was carried out of selected peptide sequences of few non UBC6s and selected UBC6s. From the MSA it was possible to predict a proposed PROSITE signature of the UBC6 family and was submitted to EBI, which is as follows:

T-[PAR]-[NS]-G-R-F-x(3)-[KTE]-[RK]-[LIV]-C-[LMS]-[ST]-[IMF]-[ST]-x(2)-H-[PK].

Figure 11.1 Multiple sequence alignment of selected ubc6s with all other yeast UBCs and their selected homologues

PROSITE signature of UBC6: T-[PAR]-[NS]-G-R-F-x(3)-[KTE]-[RK]-[LIV]-C*-[LMS]-[ST]-[IMF]-[ST]-x(2)-H-[PK]

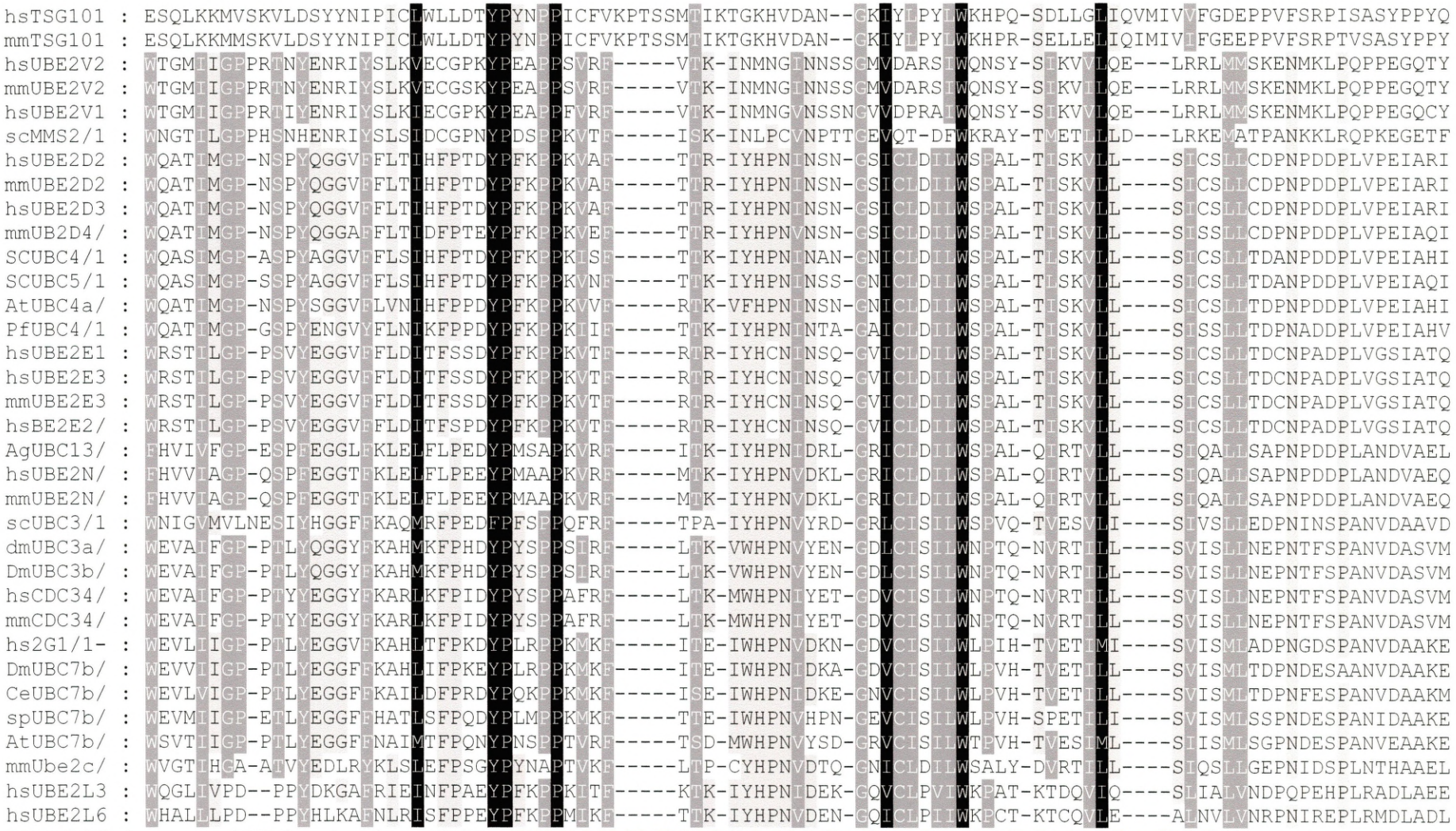


PROSITE signature of all UBCs except UBC6: [FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C*-[LIV]-x-[LIV]

** C is the active site residue*

Figure 11.1 represents the multiple sequence alignment of selected UBC6 family peptide sequences along with the other UBCs. Represented on the top and the bottom of the MSA are the prosite signatures of the UBC6 family and the prosite signatures of the UBCs other than UBC6 family. This multiple sequence alignment is a small part of the whole alignment of all 13 yeast UBCs and its homologues in those selected organisms. As the total number of sequence is 193, only a small part of the multiple sequence has been shown here highlighting the active site residue "Cysteine" and few amino acids around it which covers the PROSITE signature of UBCs.

Figure 11.2 **Multiple sequence alignment of selected UBCs other than UBC6 family, with other non-catalytic UBCs which lack the active site residue cysteine**



PROSITE signature of UBCs other than UBC6: FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C*-[LIV]-x-[LIV]

Figure 11.2 represents the multiple sequence alignment of selected UBCs other than UBC6 family, along with the other non catalytic UBCs, as TSG101 and UBE2Vs. Represented at the bottom of the MSA is again the PROSITE signature of all UBCs other than UBC6 (figure generated in GENEDOC using BLOSUM62 matrix).

PROSITE signature of all UBC's except UBC6's:-

[FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x-[LIV]*

The suggested PROSITE signature for UBC6 is:

T-[PAR]-[NS]-G-R-F-x(3)-[KTE]-[RK]-[LIV]-C-[LMS]-[ST]-[IMF]-[ST]-x(2)-H-[PK]

It can be seen from the UBCs and the UBC6 family PROSITE signatures, that as well as containing the conserved active site residue cysteine, a neighbouring [LIV] residue is also highly conserved. Other than these conserved residues, the PROSITE signature for UBCs and the UBC6 family are very different.

Using the PROSITE signature in a PHI BLAST (http://www.ncbi.nlm.nih.gov/BLAST), all known yeast UBC6 orthologues are retained. This implies that our predicted UBC6 (UBE2J) PROSITE signature is correct.

CHAPTER TWELVE

RESULTS AND DISCUSSION OF THE PHYLOGENETIC ANALYSIS OF UBCs

Figure 12.1 Phylogenetic tree of all yeast UBCs and its homologues in selected organisms generated by Phylip.

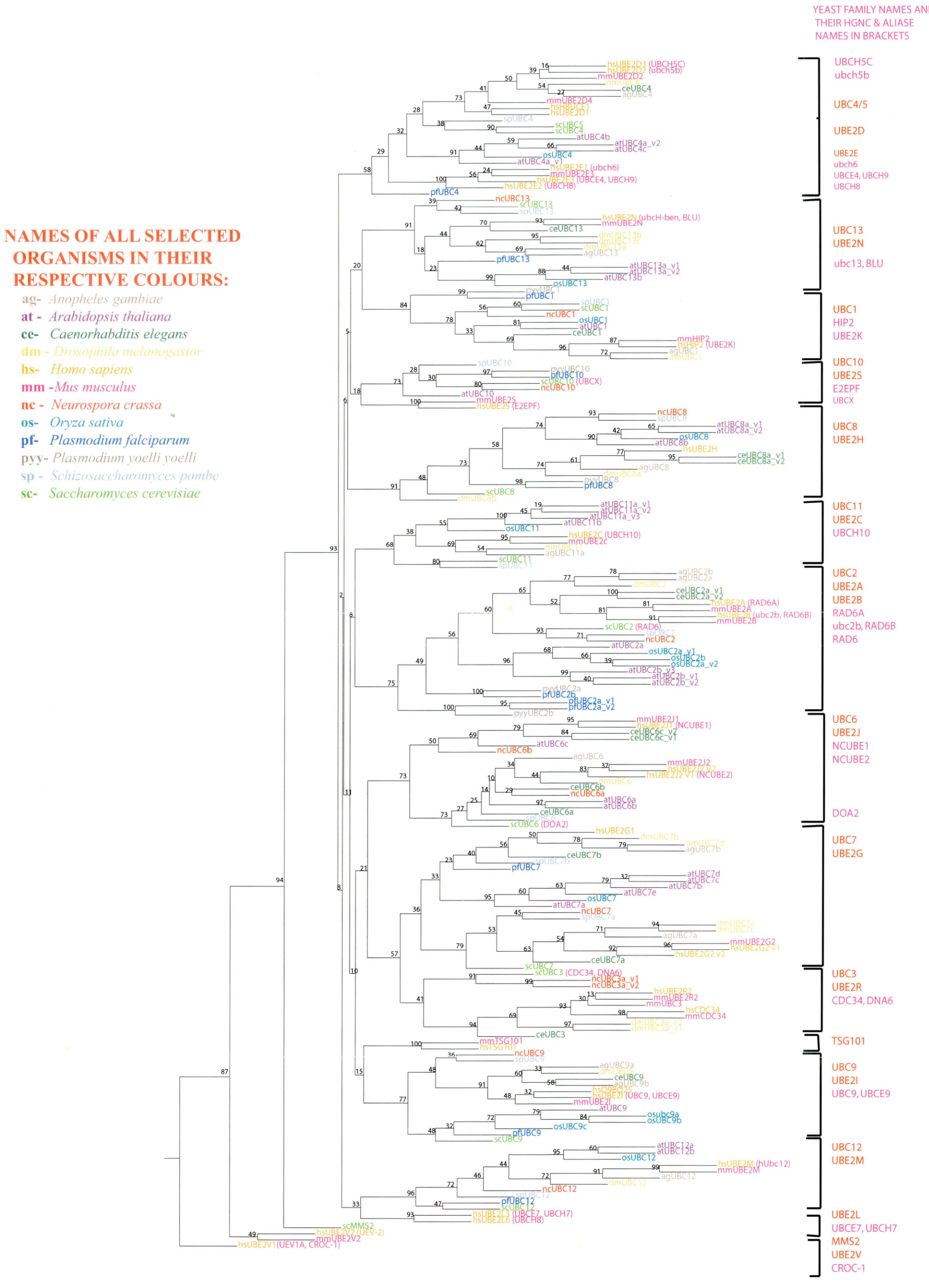


Figure 12.1 shows a Phylip 3.6 neighbour-joining phylogenetic tree for all 13 yeast UBCs and their homologues in 11 other selected organisms, whose genomes have been fully sequenced (a different colour has been assigned to each organism). In addition to the 13 yeast UBC's, the human UBE2Ls (UBE2L3 & UBE2L6) which does not have its yeast orthologue were included in the analysis. The distant UBC relative yeast MMS2, which lacks a cysteine active site (Lewis et al., 2006), and its human homologues UBE2V1 (CROC1) and UBE2V2, were also included (Rothofsky and Lin, 1997). The human and mouse TSG101 proteins (Koonin and Abagyan, 1997) which are different distant relatives of UBCs, that also lack an active site cysteine, were also included in the phylogenetic analysis.

From this figure it can clearly be seen that with the notable exception of yeast UBC4 and 5, all other yeast UBCs appear to have evolved separately, as they form their own unique evolutionary branches. These clearly distinct branches that contain probable functional orthologues, have been annotated with both the yeast UBC name and any alternate names. The accession numbers, of each of the abbreviated names of the peptide sequences of the phylogenetic tree can be found in Table-1 of the appendix section.

Figure 12.2 Phylogenetic tree of all UBCs and its homologues of selected organisms generated by Consurf

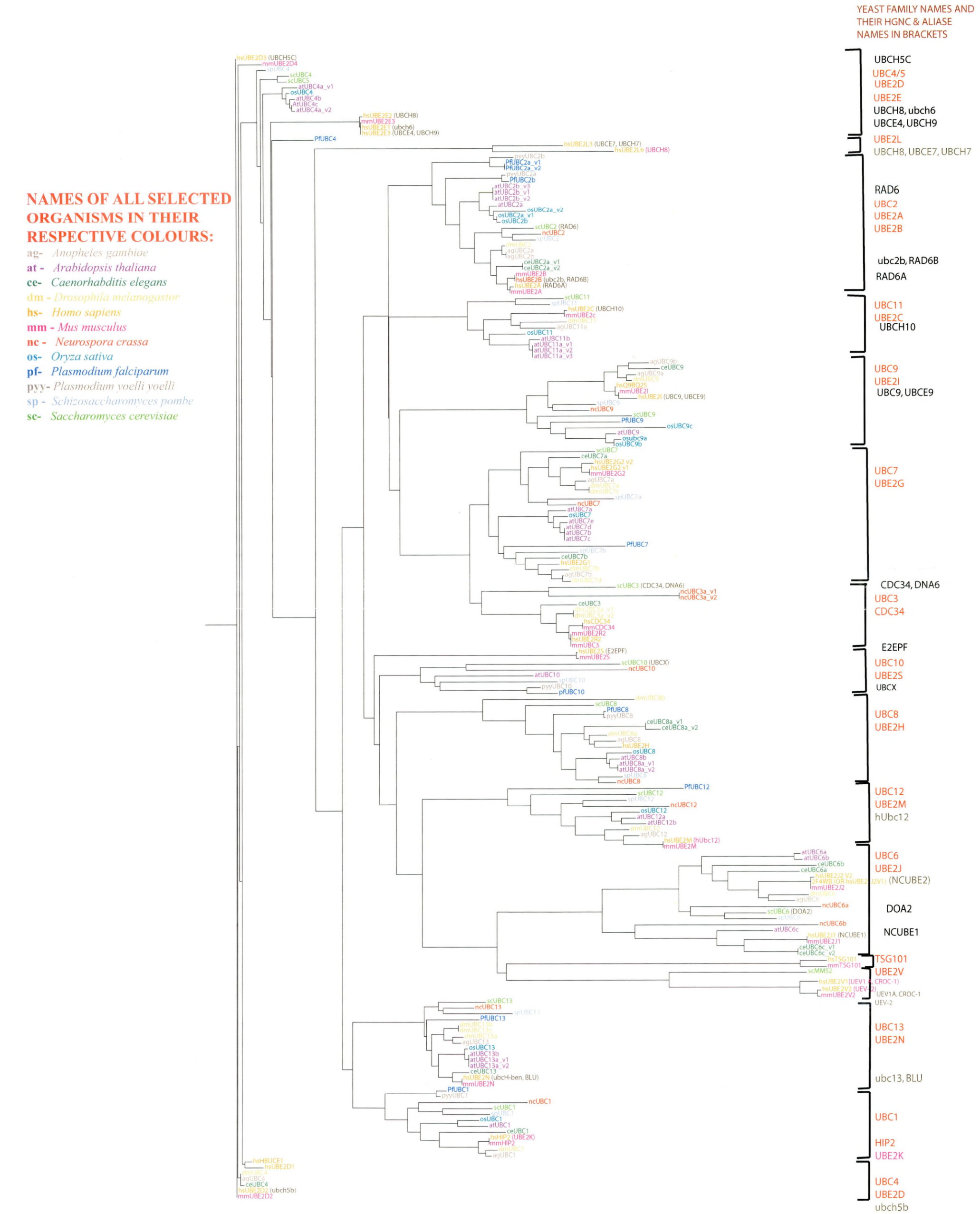


Figure 12.2 is the phylogenetic tree obtained by the Consurf analysis. The peptide sequence of the predicted structure of UBE2J2 (2F4WB) was taken as the template and all its homologues (193 peptide sequences) have been analysed in Consurf. The tree is particularly useful in that it shows structural conservations of the homologues of UBE2J2.

12.1 Results and Discussion of the phylogenetic analysis of UBCs by Phylip

As can be seen from the phylogenetic trees generated by the two different programs (Phylip 3.6 and Consurf), the tree generated by Phylip 3.6 has 15 different branches of which there are 12 branches that contain 13 out of a total of 14 yeast UBCs. Yeast UBC4 and yeast UBC5 and their respective homologues appear in the same branch of the tree presumably because of the high degree of identity (93%) between the yeast UBC4 and UBC5. The other yeast UBC-like protein, MMS2, occupies a separate branch of the tree with its human homologues, UBE2V1 and UBE2V2. The remaining branches of the tree contain the human UBE2L homologues and human & mouse TSG101 homologues respectively.

12.2. Phylogenetic analysis of peptide sequences by Consurf

A phylogenetic analysis of the same 193 peptide sequences, were also carried out using Consurf (http://conseq.bioinfo.tau.ac.il/). The phylogenetic tree generated by Consurf (shown in Figure 12.2) helped to compare and contrast the findings of the relationships of all UBCs shown by the phylogenetic tree generated by Phylip.

The phylogenetic tree generated by Consurf had also reproduced similar results as that of the Phylip tree, with the only difference that yeast UBC4 and yeast UBC5 and its homologues have been placed in two different branches. The Phylip generated tree has the yeast UBC4, yeast UBC5 and its homologues appearing in the same branch. The probable reason for the difference could be due to the difference in the method of analysis of the two programs Phylip and Consurf.

Table 12.1 Methods of phylogenetic analysis by Phylip 3.6 and Consurf

Methods of analysis	Phylip 3.6	Consurf
Tree-puzzle analysis (Gu and Zhang, 1997)	Tree-puzzle analysis is carried out to calculate the rate heterogeniety	Tree-puzzle analysis is not carried out
Bootstrapping (Felsenstein, 1985)	Bootstrapping analysis is carried out	Does not carry out the bootstrap analysis
Neighbor-joining (Felsenstein, 2004)	Neighbor-joining is used to compute the tree	Neighbor-joining is also used to compute the tree

From the comparative study of the method of analysis of the two programs, it can be said that the method of analysis used by Phylip 3.6 appears to be more rigorous than Consurf. The various steps mentioned in the Table 12.1 used by Phylip to generate the phylogenetic tree are some of the essential features which are required in the analysis of generating a statistically significant phylogenetic tree. In Consurf the method of analysis appears not to be as rigorous as Phylip, but nevertheless it had reproduced a phylogenetic tree which is very similar to that of Phylip. This might have some biological implications, but can be considered at present to be debatable. The Consurf tree as a whole, with the exception of UBC4/5 branch supports the branching of the Phylip tree.

12.3. The overall discussion of the phylogenetic trees generated by Phylip and Consurf

The Phylogenetic trees obtained by Phylip and Consurf, is a perfect example of how each UBCs have segregated themselves into each individual branches. 13 different yeast UBCs (UBC1- UBC13) have been used to find their homologues in 11 different organisms whose genomes have been fully sequenced. The 11 different organisms used are *Anopheles gambiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Neurospora crassa*, *Oryza sativa*, *Plasmodium falciparum*, *Plasmodium yoelli yoelli*, *Schizosaccharomyces pombe*.

- As appears in both the phylogenetic trees, UBC7 and UBC3 are in the adjacent branches appearing in the same node. This was previously indicated in the multiple sequence alignment as UBC7 and UBC3 have 13 extra amino acids in common which are not found in any other UBCs.
- Interestingly, the yeast MMS2 appears on its own in a separate branch, with its human orthologues UBE2Vs appearing in the adjacent branches in the Phylip generated phylogenetic tree.
- The phylogenetic tree obtained by Phylip has TSG101 in adjacent branches to UBC9, whereas in the Consurf tree TSG101 appears in adjacent branches to yeast MMS2 and UBE2Vs.
- In the phylogenetic tree generated by Phylip, UBC4/UBC5 appears in a single branch which is furthest away from the main node of the tree. The phylogenetic tree generated by Consurf has the yeast UBC4/5 and its homologues in separate

branches and are the first among all other UBC to have diverged from the main node.

- UBC13 and UBC1 appear to be in adjacent branches in both the phylogenetic trees, which illustrates that UBC13 & UBC1 may share some common evolutionary and functional similarity.
- UBC2 and UBC11 appear to be in adjacent branches in both the phylogenetic trees, which indicate some commonality in their evolution.
- UBC8 and UBC10 are in the adjacent branches of both the phylogenetic trees.
- In the phylogenetic tree obtained by Phylip, UBC6 have diverged from a common node to that of UBC7 and UBC3, whereas in the phylogenetic tree obtained by Consurf UBC9 appears to be in the adjacent branch to that of UBC7 and UBC3.

From the phylogenetic study both evolutionary and functional relationships can be estimated. These relationships are on the basis of the amino acid sequences, but more elaborate functional relationship can be assumed on the basis of 3-dimensional structures. Many of the UBC structures have been elucidated by X-ray crystallography, like *Arabidopsis thaliana* UBC1, yeast UBC4, yeast UBC7, and human UBC9. The structure of one of the human homologue of yeast UBC6 has also been found recently, i.e. UBE2J2. As the peptide sequence of the other human homologue of yeast UBC6 which is UBE2J1, is quite different from that of UBE2J2, and as its 3-dimensional structure is not known, it was essential to elucidate the structure of UBE2J1 mainly for its clinical implications.

Illustrated in the following pages are all the branches of the two phylogenetic trees, which have been cut out and inferred individually, so as to compare and contrast the relationships of each UBCs as appearing in the two phylogenetic trees.

12.3.1. The yeast UBC6 and human UBE2J orthologous phylogenetic branch

Figure 12.3 Branch of UBC6 generated by Phylip

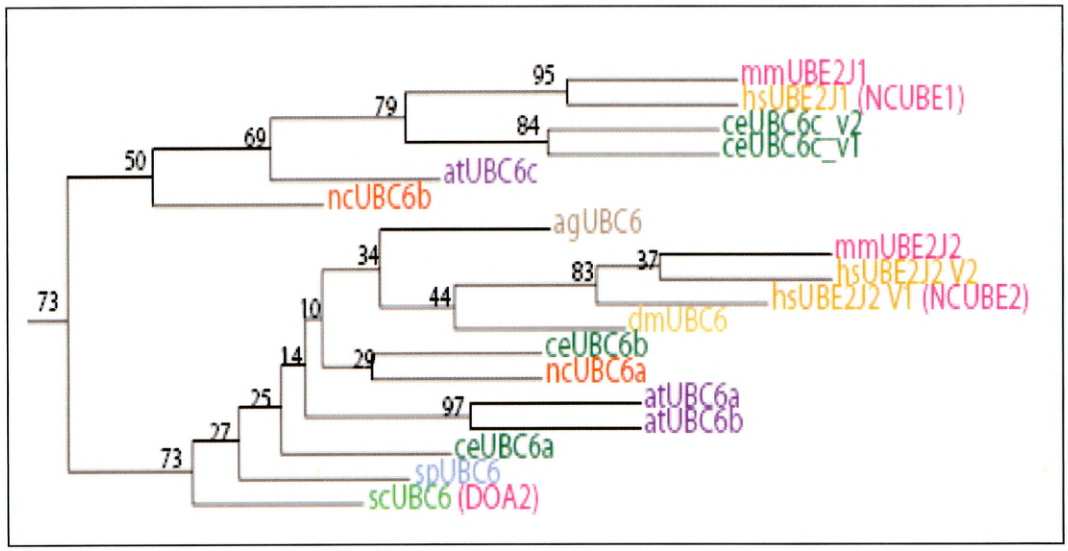
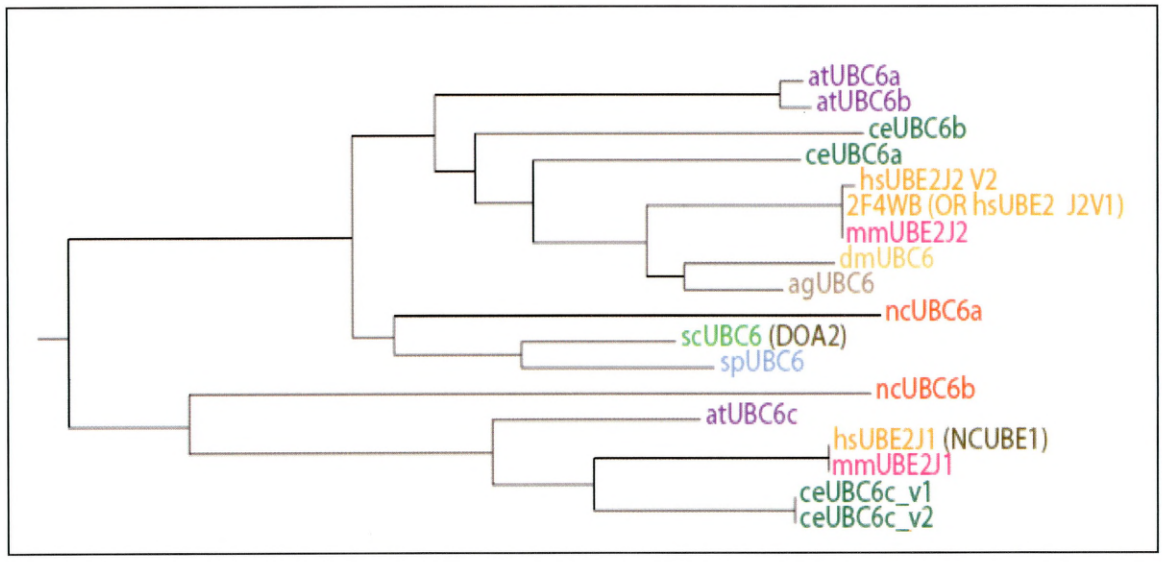


Figure 12.4 Branch of UBC6 generated by Consurf



Figures 12.3 and 12.4 illustrate sections of the two phylogenetic trees (Figure 12.1 & 12.2) the branch of yeast UBC6 and its homologues in different organisms from the Phylip and the Consurf generated trees respectively.

Table 12.2 Gene duplication in the UBC6 family

Different UBC6s	Number of genes	Different splice /start variants
Homo sapiens	(2) UBE2J1, UBE2J2	UBE2J2_v1, UBE2J2_v2
Oryza sativa	0	
Arabidopsis thaliana	(3) UBC6a, UBC6b, UBC6c	
Mus musculus	(2) UBE2J1, UBE2J2	
C.elegans	(3)UBC6a, UBC6b, UBC6c	UBC6c_v1, UBC6c_v2
P. falciparum	0	
Anopheles gambaie	1	
Neurospora crassa	(2) UBC6a, UBC6b	
S.pombe	1	
D. melanogaster	1	
S.cerevisiae	1	
P. yoelli yoelli	0	

From the Figures 12.3, 12.4, it can clearly be seen that yeast UBC6 and its homologues/orthologues have branched out very distinctly on its own, in both the phylogenetic trees generated. Only one each of yeast, *pombe, drosophila, anopheles* UBC6 gene is present in contrast to most other multicellular organisms which have undergone duplication of this gene at least once e.g. there are two homologues in *Homo sapiens and Mus musculus*, three in *A. thaliana* and three in *C. elegans*. A notable exception to this finding, are the two insects *D. melanogaster* and *A. gambiae*, which like the yeasts have only one homologue of this gene in their genomes. This latter finding may either be due to the common ancestor of insects having only one yeast UBC6 homologous gene, or that there has been complete loss of a previously duplicated gene in this ancestral species.

All the enzymes in this branch contain an ER transmembrane domain at their carboxy terminus and have or are putatively been shown to be involved in the ER associated degradation of misfolded proteins. Moreover all the enzymes on this branch have the PROSITE signature:

T-[PAR]-[NS]-G-R-F-x(3)-[KTE]-[RK]-[LIV]-C-[LMS]-[ST]-[IMF]-[ST]-x(2)-H-[PK], where the conserved C is active site cysteine, for this subfamily of UBC's.

Bootstrap values have been obtained from the phylogenetic tree generated by *Phylip*, but no bootstrap values have been obtained by the other tree that had been generated by *Consurf*. The reason for not having the bootstrap values for the Consurf generated phylogenetic tree is that Consurf generated only the phylogenetic tree and plotted the conservation of the amino acids of UBCs onto the structure of UBE2J1. It does not estimate the bootstrap values.

There are bootstrap values in the tree generated by Phylip, ranging from very low of 10% to a high of 97%. Bootstrap values of 90% and above considered as very confident assumptions that the branching has been evolved and the rate of appearance of these same species in the same branch in any circumstance is very high. But bootstrap values less than 50% are not considered as confident assumptions. Irrespective of the bootstrap values, as yeast UBC6 and all its homologues in all different species have segregated themselves into one single branch apart from all other UBCs, is a clear indication that UBC6s are evolutionarily and functionally distinct from all other UBCs.

What is more interesting in the Consurf generated phylogenetic tree, is that the other UBC homologues which lack the UBC active site cysteine i.e. TSG101 and UBE2V1 (CROC-1) & UBE2V2 are in the adjacent branches of UBC6. This shows that TSG101, though it lacks the active site, could be very much related to the UBC6 evolutionarily and functionally. This also indicates that there might have been a common ancestor to these genes and might have diverged in the time of evolution. There are two types of evolution, convergent evolution and divergent evolution. Divergent evolution is two or more related species becoming more and more dissimilar. Convergent evolution on the other hand is unrelated species becoming more and more similar to each other in the course of evolution. Here the appearance of TSG101 and UBE2V1 (CROC-1) indicates a convergent evolution.

12.3.2. Yeast UBC9 and its homologues

Figure 12.5 UBC9 branch generated by Phylip

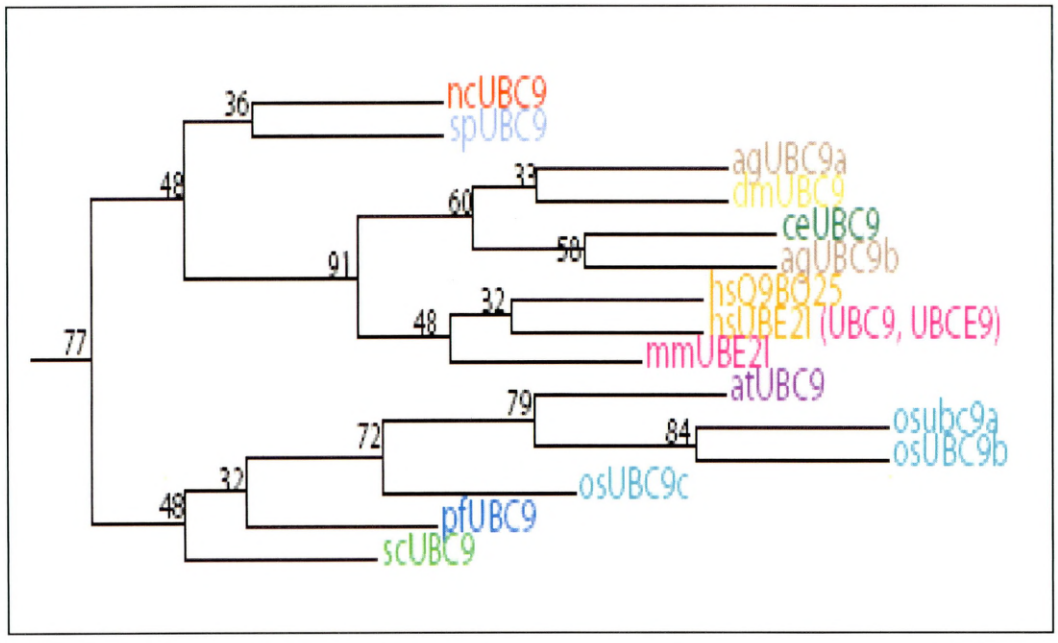
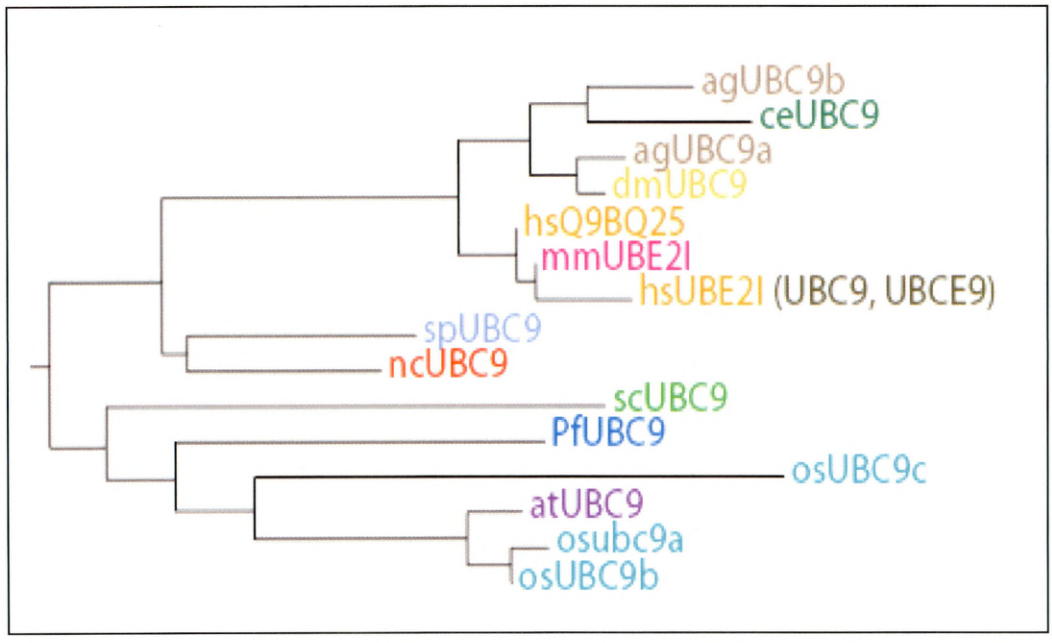


Figure 12.6 UBC9 branch generated by Consurf



Figures 12.5 and 12.6 illustrates a section taken from the Phylip and the Consurf generated trees respectively, highlighting the yeast UBC9 and its homologues.

As evident from the above figures that yeast UBC9 and all its homologues have also appeared in a single branch. Both the phylogenetic trees look very similar, which gives a confidence of how well UBC9 has segregated into a single branch, distinct from all other UBCs.

Table 12.3 Gene duplication in the UBC9 family

Different UBC9s	Number of genes	Different splice /start variants
Homo sapiens	(2), UBE2I, Q9BQ25	
Oryza sativa	(3), UBC9a, Ubc9b, UBC9c	
Arabidopsis thaliana	1	
Mouse	1	
C.elegans	1	
Plasmodium falciparum	1	
Anopheles gambaie	(2), UBC9a, UBC9b	
Neurospora crassa	1	
S.pombe	1	
D. melanogaster	1	
S.cerevisiae	1	
P. yoelli yoelli	0	

From Table 12.3 it can be clearly seen that other than *Homo sapiens, Oryza sativa and Anopheles gambaie*, all other organisms appear to have appeared only once. There are three orthologues of *Oryza sativa*, two of *Homo sapiens*, and two in *Anopheles gambaie.*

12.3.3. Yeast UBC11 and its homologues

Figure 12.7 UBC11 branch generated by Phylip

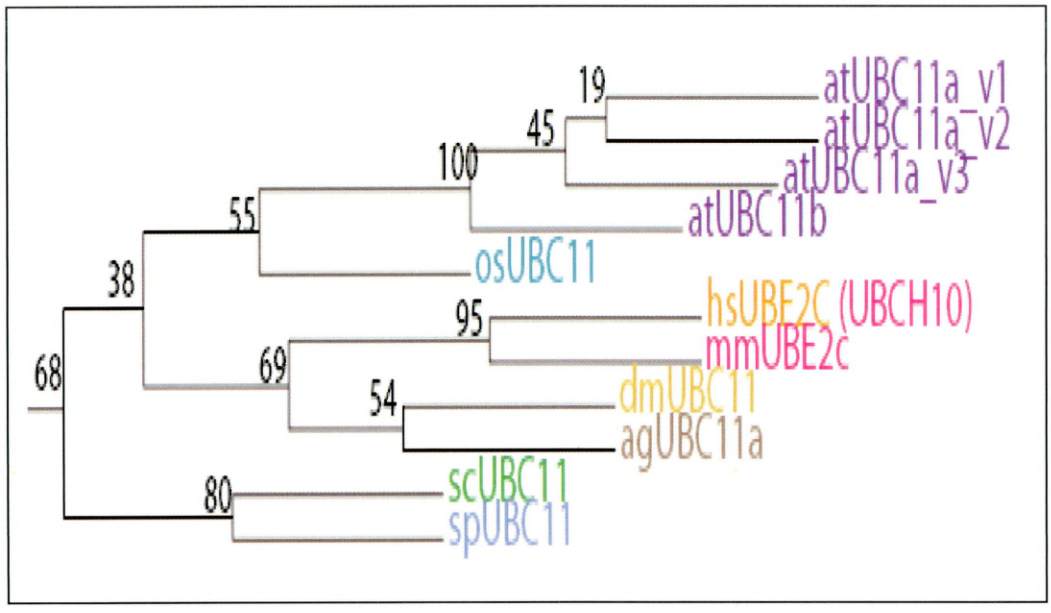
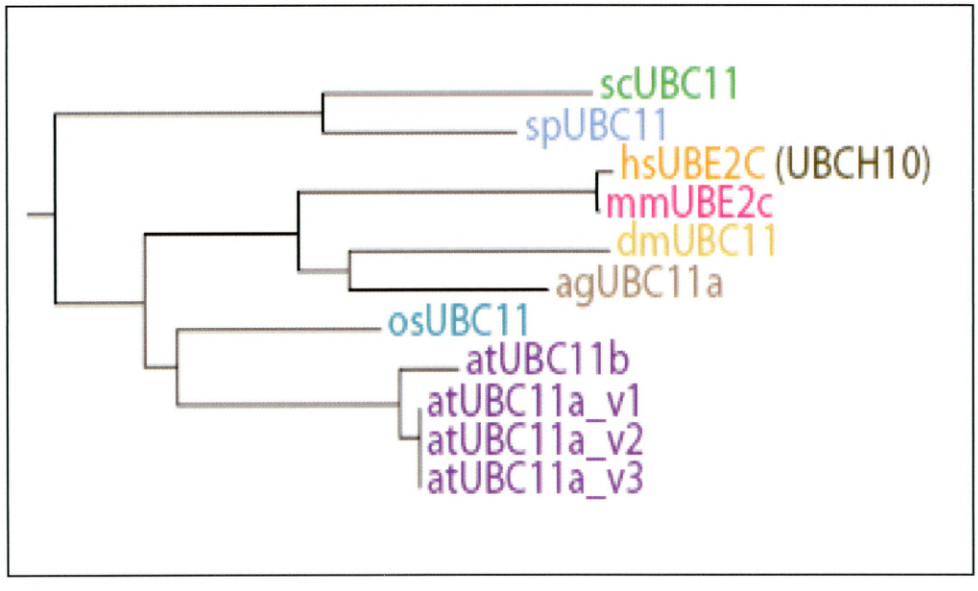


Figure 12.8 UBC11 branch generated by Consurf



Figures 12.7 and 12.8, highlights a section taken from the main Phylip and the Consurf generated trees respectively, which illustrates the branching of yeast UBC11 and its homologues.

The branches of this section of the phylogenetic trees above, shows the segregation of yeast UBC11 and its homologues from all other UBCs. They have evolved in the same branch with bootstrap values in the range of 100%, 95%, which are perfect confirmations of the assumption that the yeast UBC11 and its homologues have appeared in the same branch. As it appears in the UBC11, *Arabidopsis* have evolved more than once.

Table12.4 Gene duplication in the UBC11 family

Different UBC11s	Number of genes	Different splice /start variants
Homo sapiens	1	
Oryza sativa	1	
Arabidopsis thaliana	(2), UBC11a, UBC11b	UBC11a_v1, UBC11a_v2, UBC11a_v3.
Mouse	1	
C.elegans	0	
Plasmodium falciparum	0	
Anopheles gambaie	1	
Neurospora crassa	0	
S.pombe	1	
D. melanogaster	1	
S.cerevisiae	1	
P. yoelli yoelli	0	

From Table 12.4 it is evident that with the exception of *Arabidopsis*, all other organisms have appeared only once in UBC11. Only *Arabidopsis* have duplicated as it appears to have two genes with three splice variants of one of the two genes as can be seen in the table above.

12.3.4. Yeast UBC2 and its homologues

Figure 12.9 UBC2 branch generated by Phylip

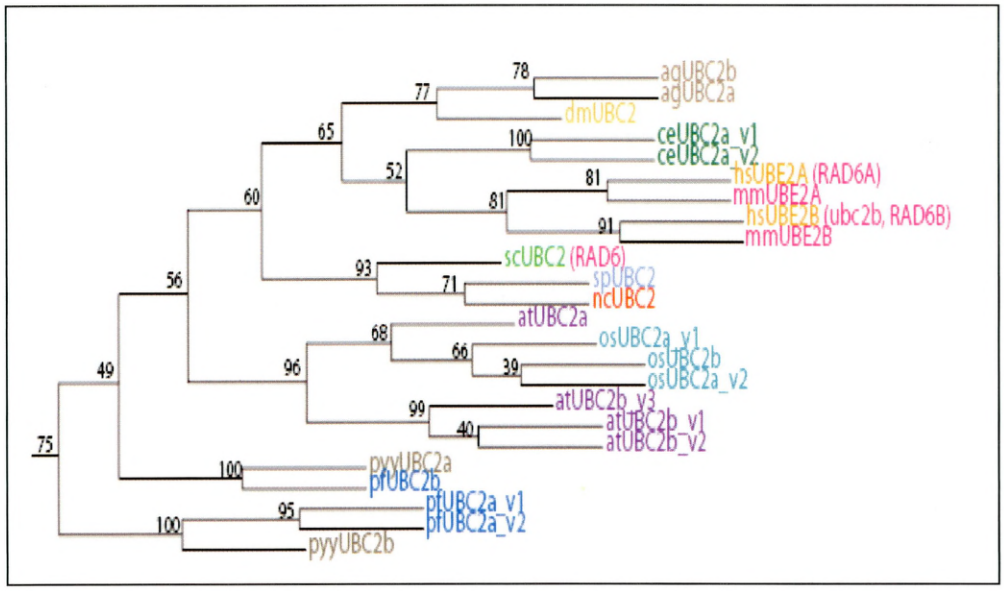


Figure 12.9 shows the phylogenetic branching of yeast UBC2 and its homologues in different organisms generated by Phylip.

Figure12.10 UBC2 branch generated by Consurf

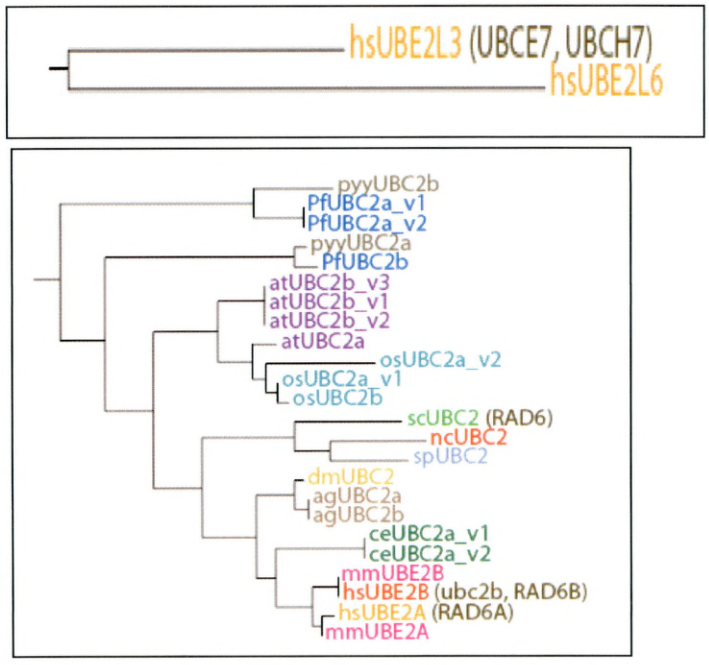


Figure 12.10 illustrates two adjacent branches of the main phylogenetic tree generated by Consurf, one of which is the human orthologues UBE2Ls and the other is the branch of yeast UBC2 and its homologues. The reason for illustrating the two branches here is to highlight that that the UBE2L family is closer to the UBC2 family.

The branches illustrated in Figure 12.10 of the phylogenetic tree consists of the branch of human orthologues UBE2L3 & UBE2L6 which is adjacent to the branch of yeast UBC2 and its homologues. The phylogenetic tree obtained by Phylip and Consurf has the yeast UBC2 and its orthologues in a single branch. The UBE2L branch of the Consurf tree has been shown here as it appears in the adjacent branch to UBC2 in the Consurf generated tree. In the phylip tree the UBE2L branch appear adjacent to UBC12 family. Although the Phylip generated tree is more rigorous than the Consurf tree, the bootstrap value in the phylip tree is very poor (33%). This low bootstrap value indicates that it is not confirmed that the UBE2L branch and the UBC12 family branch are quite close to each other in evolution. It could be that the Consurf result may be right by illustrating the UBE2Ls being closer to the UBC2 family.

Table12.5 Gene duplication in the UBC2 family

Different UBC2s	Number of genes	Different splice /start variants
Homo sapiens	2, UBE2A, UBE2B	
Oryza sativa	2, UBC2a, UBC2b	UBC2a_v1, UBC2a_v2
Arabidopsis thaliana	2, UBC2a,UBC2b	UBC2b_v1, UBC2b_v2, UBC2b_v3 are splice variants or polymorphic variant.
Mouse	2, UBE2A, UBE2B	
C.elegans	1, UBC2a	UBC2a_v1, UBC2a_v2
Plasmodium falciparum	2, UBC2a, UBC2b	UBC2a_v1, UBC2a_v2
Anopheles gambaie	2, UBC2a, UBC2b	
Neurospora crassa	1	
S.pombe	1	
D. melanogaster	1	
S.cerevisiae	1	
P. yoelli yoelli	2, UBC2a, UBC2b	

It can be seen that with the exception of yeast, *Neurospora crassa*, *S. pombe*, and *D. melanogaster*, all organisms have duplicated in the UBC2 family. *C. elegans* has also one gene, but it has two splice variants.

12.3.5. The yeast UBC7/ human UBE2G branch and its evolutionary relationship to the yeast UBC3/human CDC34 branch.

Figure 12.11 UBC7 and UBC3 branch generated by Phylip

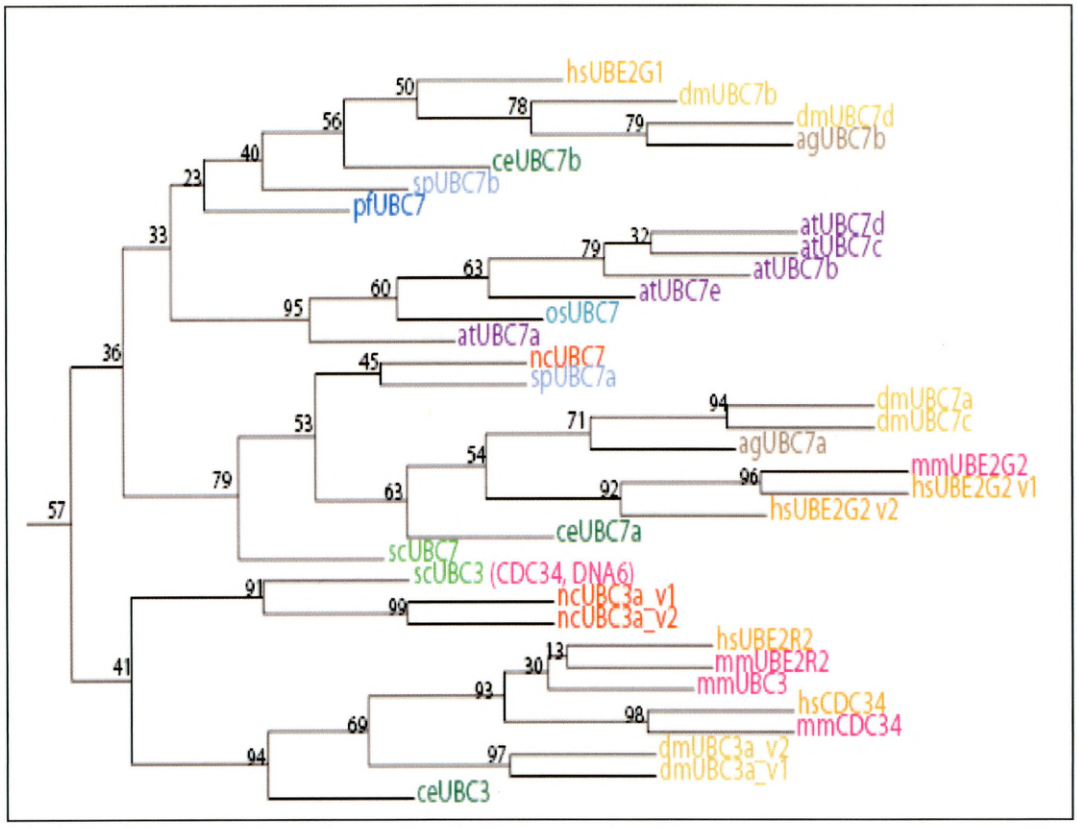


Figure 12.11 is the phylogenetic branch cut out from the main phylogenetic tree generated by Phylip, illustrating the close evolutionary relationship of yeast/human UBC7/UBE2G and their related orthologues to yeast/human UBC3/CDC34.

Figure 12.12 UBC7 and UBC3 branch generated by Consurf

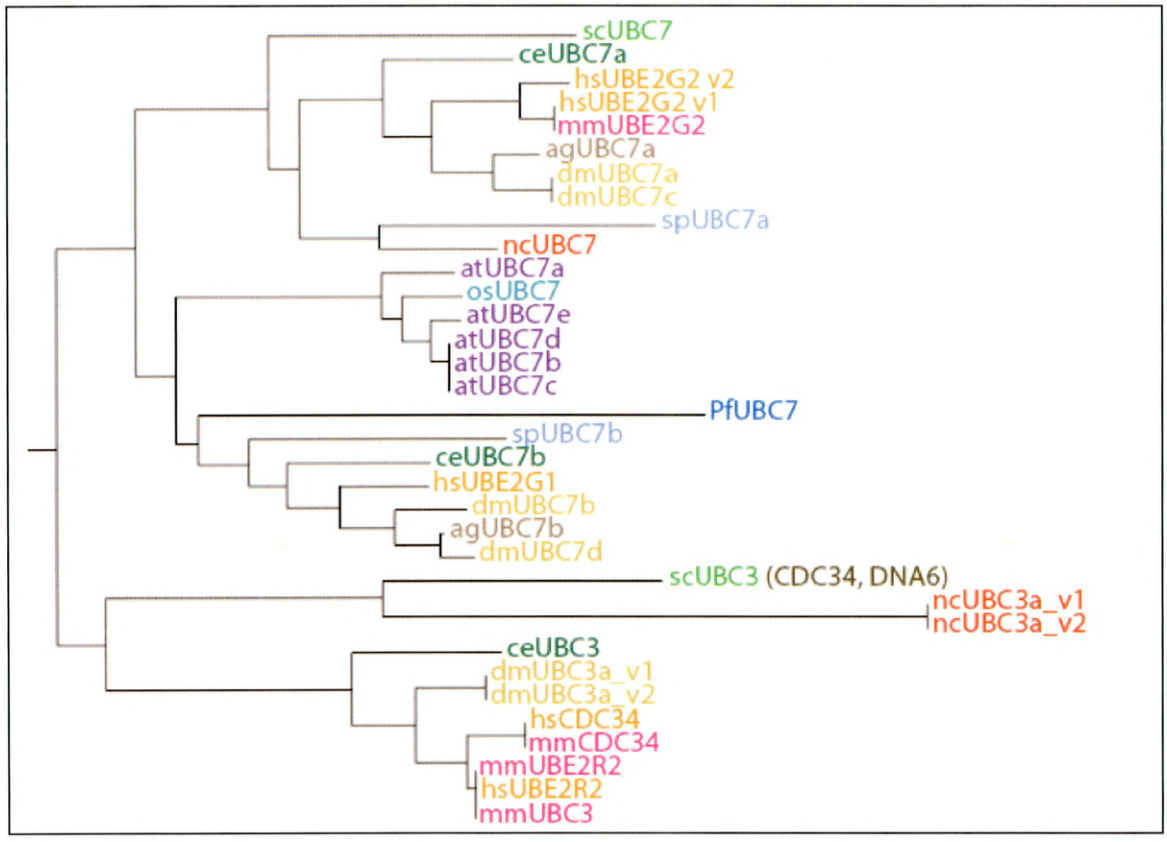


Figure 12.12 is the phylogenetic branch cut out from the main phylogenetic tree generated by Consurf, illustrating the close evolutionary relationship of yeast/human UBC7/UBE2G and their related orthologues to yeast/human UBC3/CDC34.

As can be seen from the multiple sequence alignment of all UBCs, all homologous members of yeast UBC7 and yeast UBC3 including human CDC34 have a unique, 13 amino acid insert after the conserved UBC active site residue cysteine. This 13 amino acid sequence is highly conserved among the homologues of these enzymes making likely that they all have evolved from a single common ancestral gene. This is further confirmed from the phylogenetic trees, as the yeast UBC7 and UBC3 homologues appear to have diverged from the same evolutionary node and are in the adjacent branches. This evolutionary branch bifurcation into two distinct sub-branches, one containing the yeast UBC7 homologues, which are involved in ERAD and the other which includes the human CDC34, which is involved in cell cycle control, is a clear indication that UBC7 and UBC3 are evolutionarily and functionally related.

Table 12.6 Gene duplication in the UBC7 family

Different UBC7s	Number of genes	Different splice /start variants
Homo sapiens	2, UBE2G1, UBE2G2	UBE2G2_v1, UBE2G2_v2
Oryza sativa	1	
Arabidopsis thaliana	5	
Mouse	1	
C.elegans	2	
Plasmodium falciparum	1	
Anopheles gambaie	2	
Neurospora crassa	1	
S.pombe	2	
D. melanogaster	4	
S.cerevisiae	1	
P. yoelli yoelli	0	

From Table 12.6 it can be seen that there are 5 genes present in *Arabidopsis*, four genes in *Drosophila*. All other organisms have duplicated twice, with the exception of *S. cerevisiae, N. crassa, O. sativa and P. falciparum* that appear only once.

Table 12.7 Gene duplication in the UBC3 family

Different UBC3s	Number of genes	Different splice /start variants
Homo sapiens	2	
Oryza sativa	0	
Arabidopsis thaliana	0	
Mouse	3	
C.elegans	1	
Plasmodium falciparum	0	
Anopheles gambaie	0	
Neurospora crassa	1, UBC3a	UBC3a_v1, UBC3a_v2
S.pombe	0	
D. melanogaster	1, UBC3a	UBC3a_v1, UBC3a_v2, splice or promoter variants.
S.cerevisiae	1	
P. yoelli yoelli	0	

From Table 12.7 it can be seen that in UBC3, there are no *O. sativa, A. thaliana, P. falciparum, A. gambaie, S. pombe and P. yoelli yoelli.* Yeast, *N. crassa, and C. elegans* have appeared once, and only *Homo sapiens* and *Mus musculus* have duplicated.

12.3.6. Yeast UBC10 and its homologues

Figure 12.13 UBC10 branch generated by Phylip

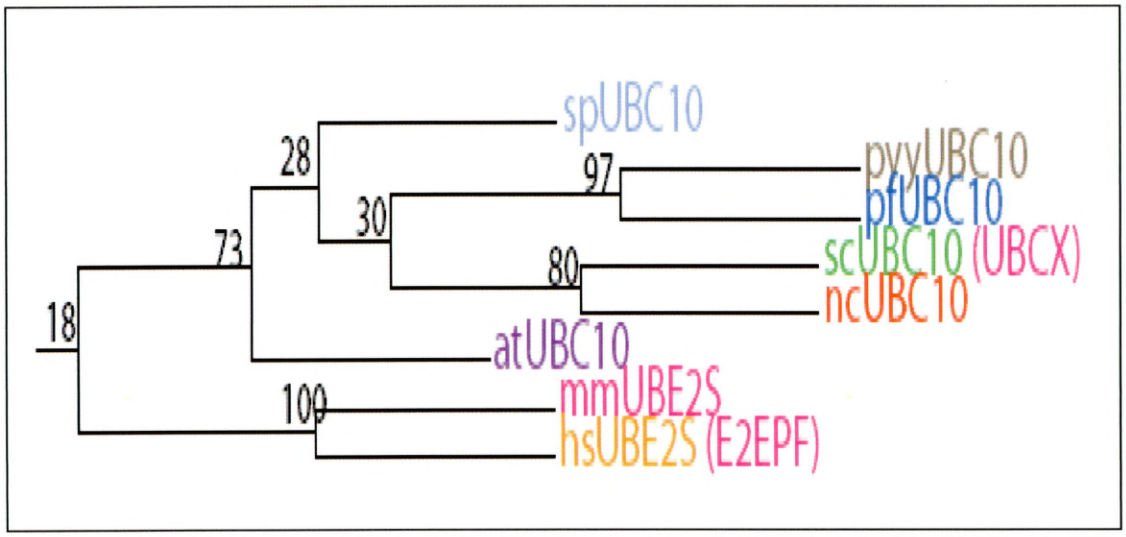
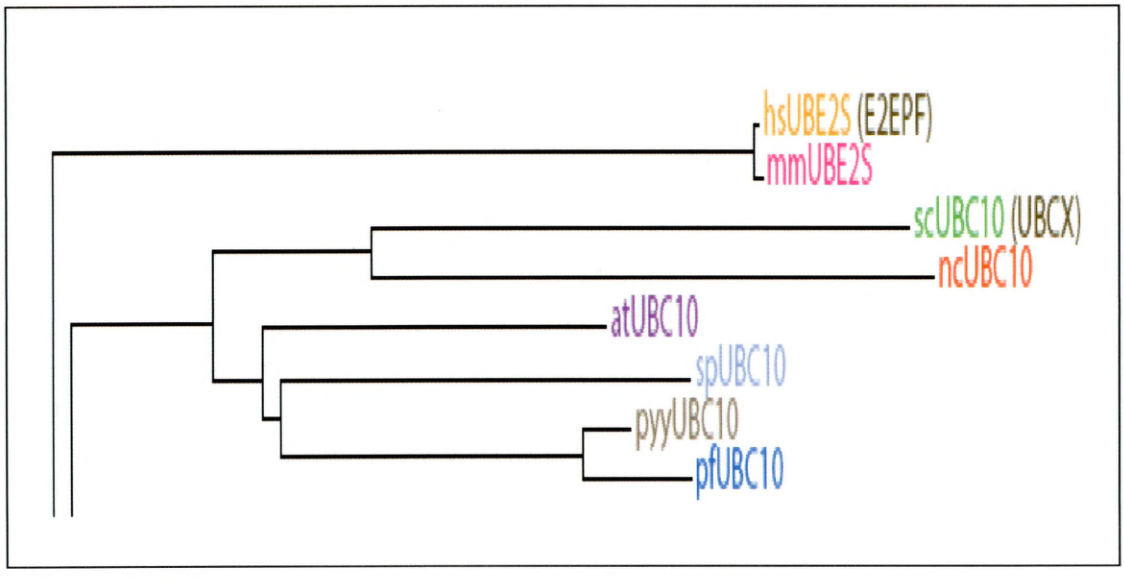


Figure12.14 UBC10 branch generated by Consurf



Figures 12.13 and 12.14 are the sections of the main phylogenetic trees generated by Phylip and Consurf respectively, showing the branching of yeast UBC10 and its homologues in different species.

Table 12.8 Gene duplication in the UBC10 family

Different UBC10s & UBC1s	Number of genes	Different splice /start variants
Homo sapiens	1	
Oryza sativa	0	
Arabidopsis thaliana	1	
Mouse	1	
C.elegans	0	
Plasmodium falciparum	1	
Anopheles gambaie	0	
Neurospora crassa	1	
S.pombe	1	
D. melanogaster	0	
S.cerevisiae	1	
P. yoelli yoelli	1	

It can be seen from the table above that all organisms have appeared only once in the UBC10 family with no duplication, with the exception of *O. sativa, C. elegans, A. gambaie, and D. melanogaste*r which have not appeared even once.

12.3.7. Yeast UBC12 and its homologues

Figure 12.15 UBC12 branch generated by Phylip

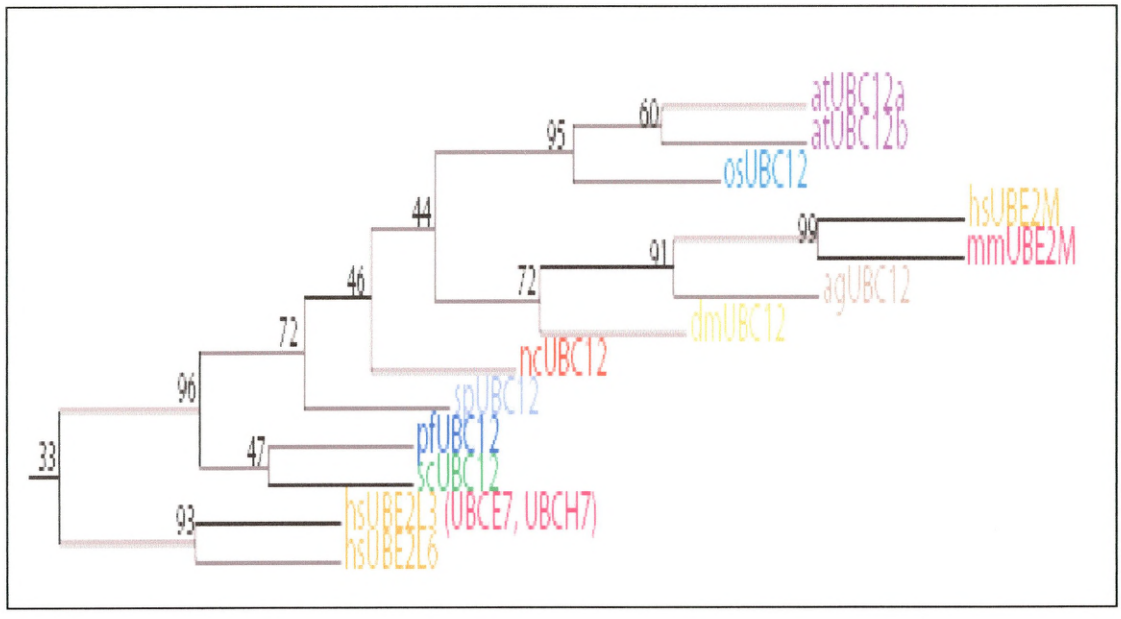
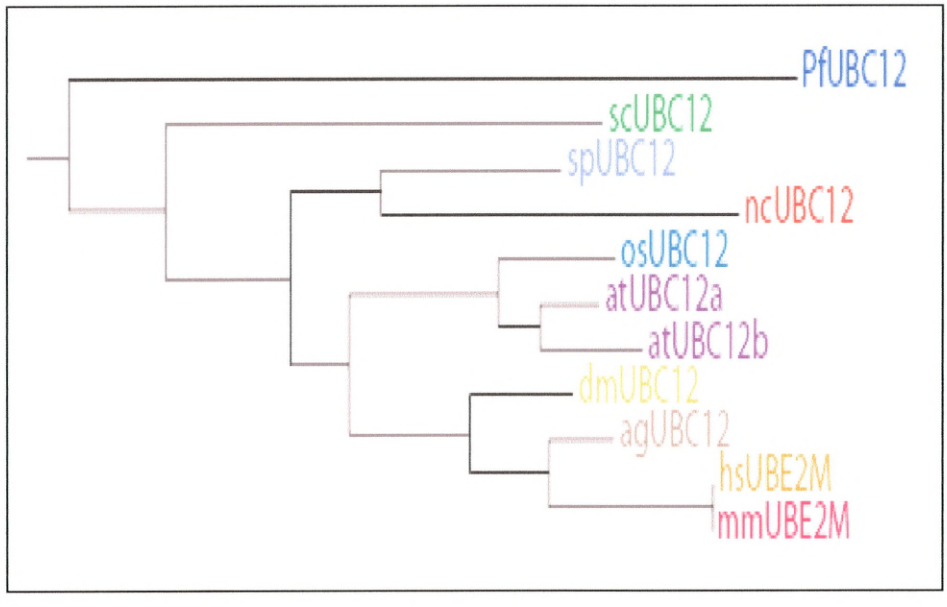


Figure 12.15 is a branch cut out from the main phylogenetic tree generated by Phylip of yeast UBC12 and its homologues along with UBE2Ls

Figure 12.16 UBC12 branch generated by Consurf



Figures 12.16 show the branch portion of the yeast UBC12 and its homologues cut out from the main tree generated by Consurf.

The difference in the phylogenetic branching of the two generated trees is that the UBE2Ls appear to have diverged from the same node as of yeast UBC12 and its homologues in the tree generated by Phylip, whereas they are distant apart in the tree generated by Consurf.

Table 12.9 Gene duplication in the UBC12 family

Different UBC12s	Number of genes	Different splice /start variants
Homo sapiens	1	
Oryza sativa	1	
Arabidopsis thaliana	2, UBC12a, UBC12b	
Mouse	1	
C.elegans	0	
Plasmodium falciparum	1	
Anopheles gambaie	1	
Neurospora crassa	1	
S.pombe	1	
D. melanogaster	1	
S.cerevisiae	1	
P. yoelli yoelli	0	

From Table 12.9, it can be seen that there are not much duplication of genes of the different organisms, with the exception of *A. thaliana* where it has duplicated.

12.3.8. Yeast UBC8 and its homologues

Figure 12.17 UBC8 branch generated by Phylip

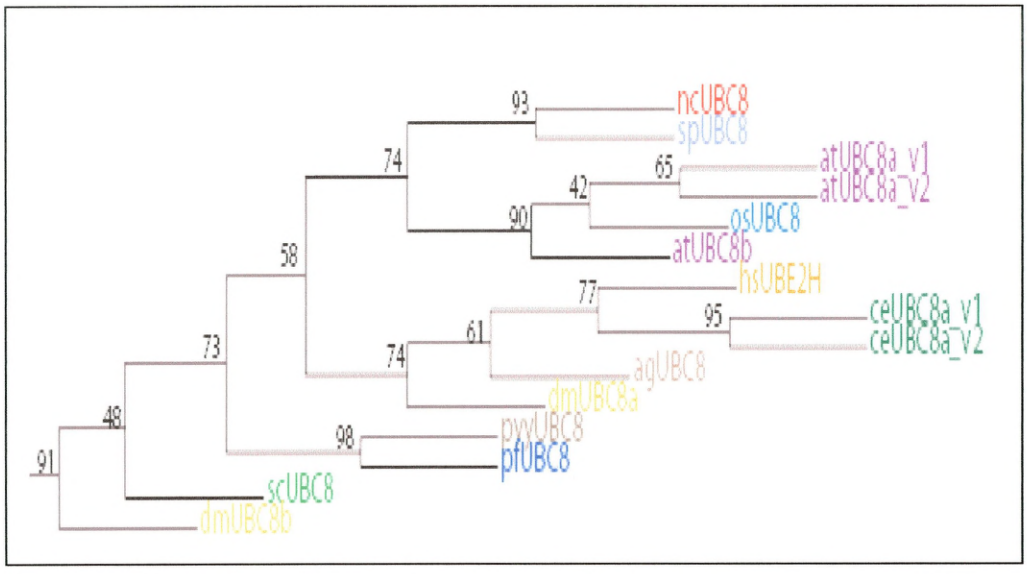
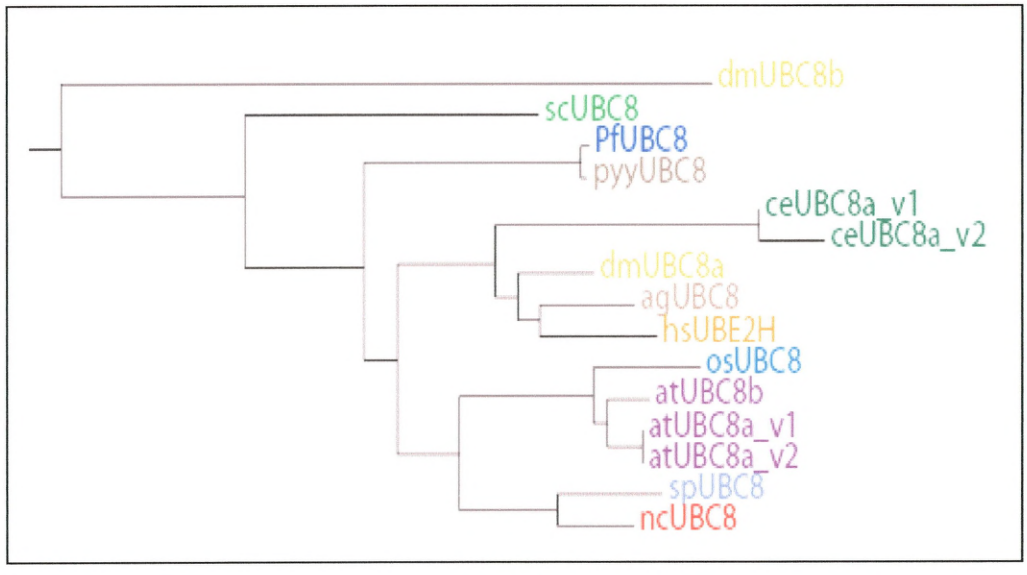


Figure 12.18 UBC8 branch generated by Consurf



Figures 12.17 & 12.18 illustrate the phylogenetic sub-branch of yeast UBC8 and its homologues in different organisms, generated by Phylip and Consurf respectively.

It can be seen that the branches of yeast UBC8 and its homologues in the two phylogenetic trees have quite distinctly segregated into a single branch. Both the phylogenetic branched trees confirm this fact that yeast UBC8 and its homologues appear to have a common evolutionary and structural similarity to each other.

Table 12.10 Gene duplication in the UBC8 family

Different UBC8s	Number of genes	Different splice /start variants
Homo sapiens	1	
Oryza sativa	1	
Arabidopsis thaliana	2, UBC8a, UBC8b	UBC8a_v1, UBC8a_v2
Mouse	0	
C.elegans	1	UBC8a_v1, UBC8_v2
Plasmodium falciparum	1	
Anopheles gambaie	1	
Neurospora crassa	1	
S.pombe	1	
D. melanogaster	2	
S.cerevisiae	1	
P. yoelli yoelli	1	

From Table 12.10 above it can be seen that only *Arabidopsis thaliana* and *Drosophila melanogaster* has duplicated in UBC8, all other organisms have appeared only once. *Mus musculus* has not appeared even once.

12.3.9. Yeast UBC13 and its homologues

Figure 12.19 UBC13 branch generated by Phylip

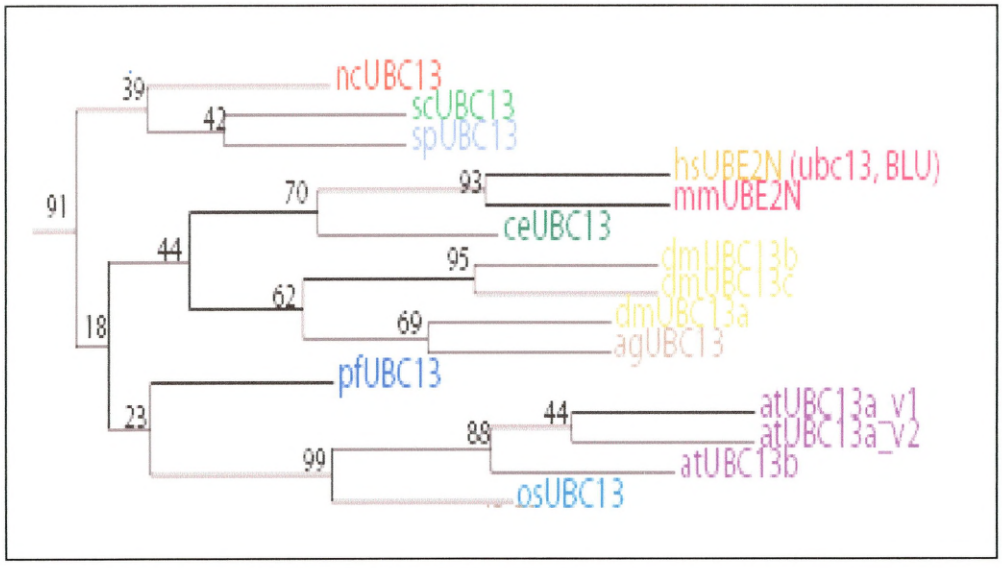


Figure 12.19 illustrates a section of the whole phylogenetic tree generated by Phylip illustrating the branch of yeast UBC13 and its homologues.

Figure 12.20 UBC13 branch generated by Consurf

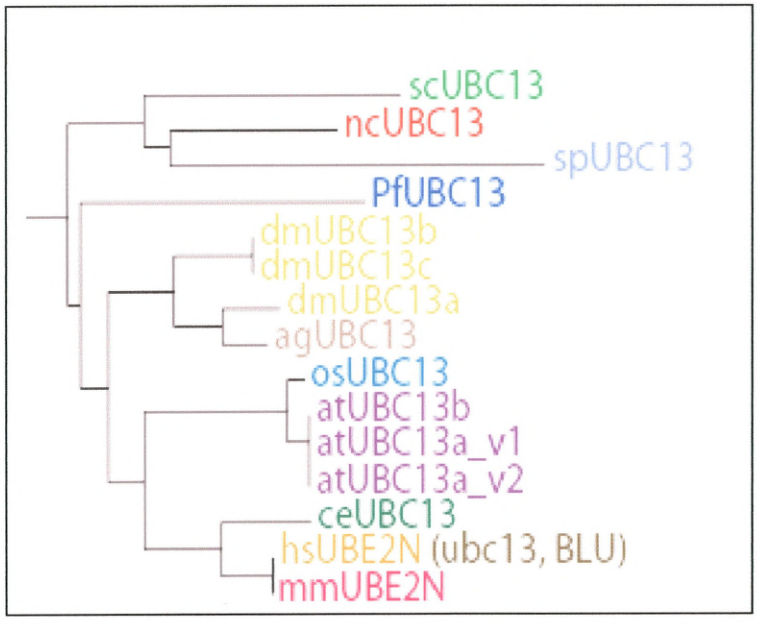


Figure 12.20 illustrates a section of the whole phylogenetic tree generated by Consurf illustrating the branch of yeast UBC13 and its homologues.

The phylogenetic trees generated by Phylip and Consurf, have very well segregated yeast UBC13 and its homologues into individual branches. This is evident that UBC13 family have evolved independently and have a function distinct from all other UBCs.

Table 12.11 Gene duplication in the UBC13 family

Different UBC13s	Number of genes	Different splice /start variants
Homo sapiens	1	
Oryza sativa	1	
Arabidopsis thaliana	2, UBC13a, UBC13b	UBC13a_v1, UBC13a_v2
Mouse	1	
C.elegans	1	
Plasmodium falciparum	1	
Anopheles gambaie	1	
Neurospora crassa	1	
S.pombe	1	
D. melanogaster	3	
S.cerevisiae	1	
P. yoelli yoelli	0	

The appearance of three peptide sequences of drosophila UBC13 and *Arabidopsis* UBC13 in the UBC13 branch is a clear indication that all peptide sequences of each of *drosophila* and *Arabidopsis* UBC13 have duplicated. There is one peptide sequence each of *S.cerevisiae, S. pombe, N. crassa, O. sativa and M. musculus*, which states that they have evolved only once and have not multiplied in the course of evolution.

12.3.10. Yeast UBC4/UBC5 and its homologues

Figure 12.21 UBC4/5 branch generated by Phylip

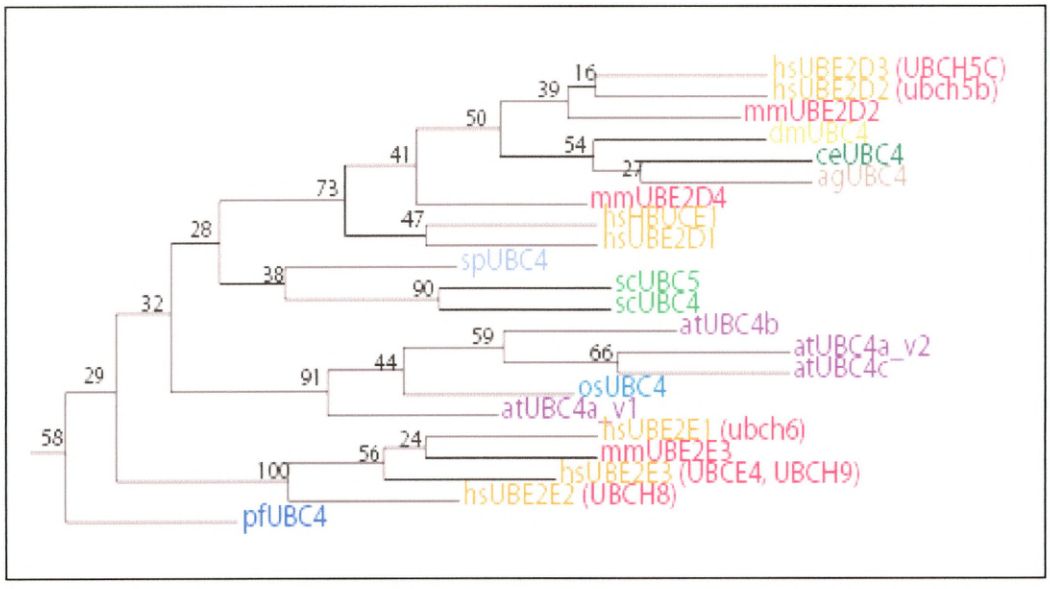


Figure 12.22 UBC4/5 branch generated by Consurf

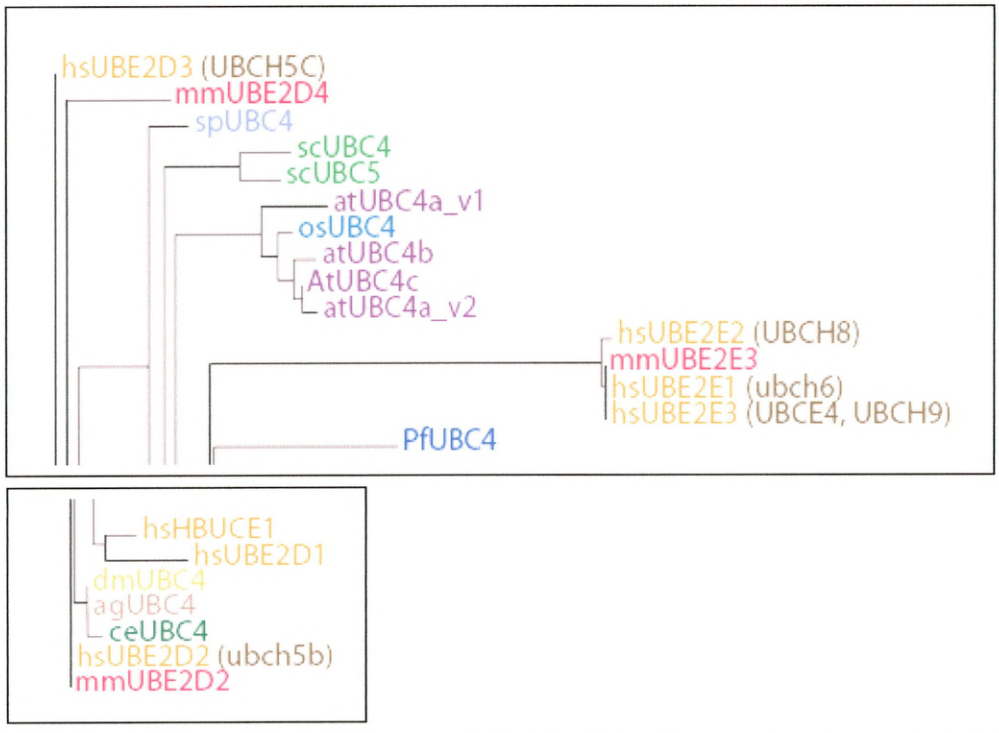


Figure 12.21 represent a section of the whole phylogenetic tree generated by Phylip of yeast UBC4/5 and its homologues in different species. Figure 12.22 represents the sections of the whole phylogenetic tree generated by Consurf of yeast UBC4/5 and its homologues in different species. Here in Consurf the yeast UBC4/5 and its homologues have been segregated into different branches.

Yeast UBC4 and yeast UBC5 peptide sequences are quite similar to each other (93% identity & 97% similarity) and its homologues are the same. There are peptide sequences of organisms like *Homo sapiens*, *Mus musculus* and *Arabidopsis thaliana*, which appear more than once. All the UBC4/5 homologues have appeared in the same phylogenetic branch in the tree generated by Phylip, but the tree generated by Consurf has the yeast UBC4/5 and its homologues in separate branches.

Table 12.12 Gene duplication in the UBC4/5 family

Different UBC4/5s	Number of genes	Different splice /start variants
Homo sapiens	7	
Oryza sativa	1	
Arabidopsis thaliana	3, UBC4a, UBC4b, UBC4c	UBC4a_v1, UBC4a_v2
Mouse	3	
C.elegans	1	
Plasmodium falciparum	1	
Anopheles gambaie	1	
Neurospora crassa	0	
S.pombe	1	
D. melanogaster	1	
S.cerevisiae	1	
P. yoelli yoelli	0	

Table 12.12 above shows that there are seven genes in *Homo sapiens*, three genes of *A. thaliana*, three of *M. musculus*, but all other organisms have appeared only once.

12.3.11. Yeast UBC1and its homologues

Figure 12.23 UBC1 branch generated by Phylip

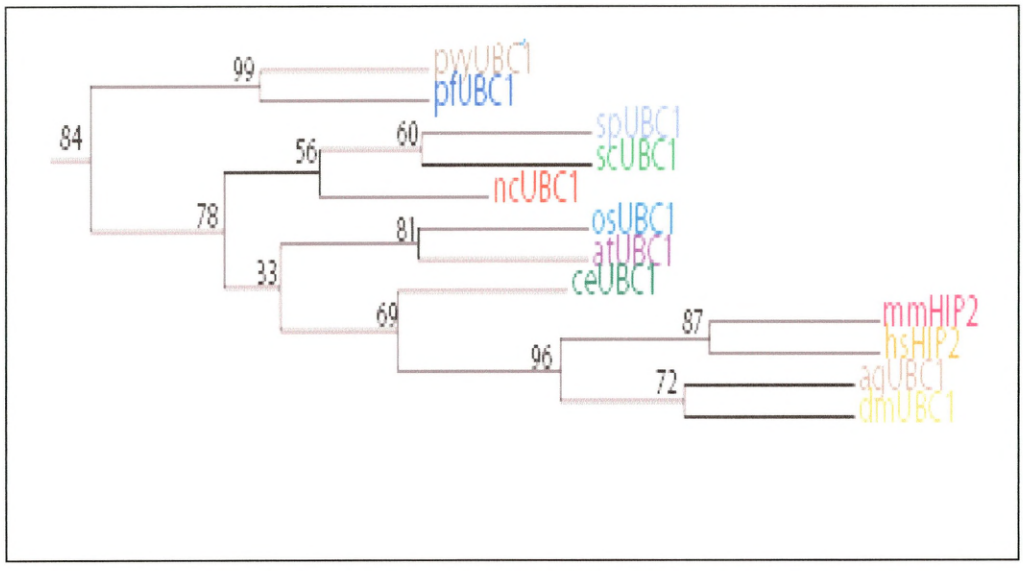


Figure 12.24 UBC1 branch generated by Consurf

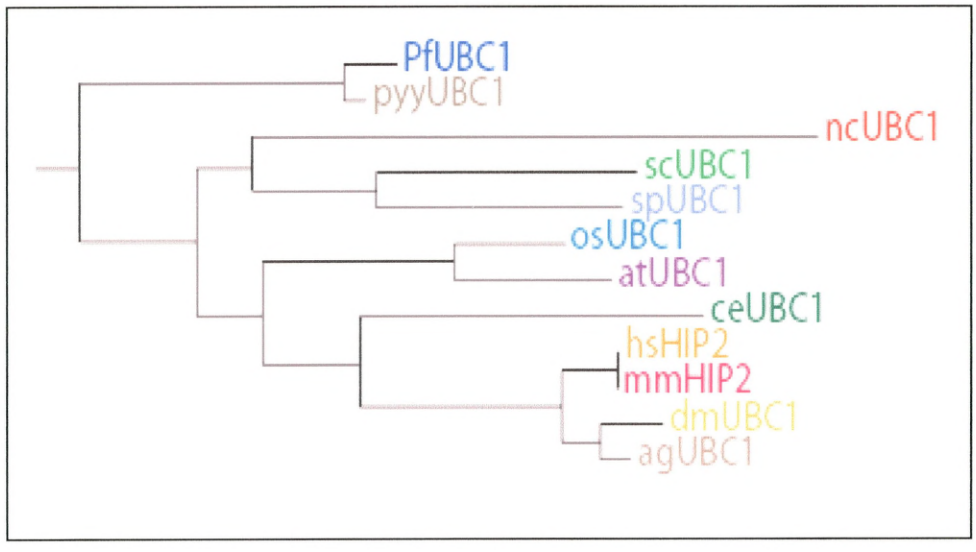


Figure 12.23 represent a section of the whole phylogenetic tree generated by Phylip the branch of yeast UBC1 and its homologues in different species, whereas Figure 12.24 is a section of the phylogenetic trees generated by Consurf the branch of yeast UBC1 and its homologues.

As can be seen from the phylogenetic sub-branches from the two trees generated by Phylip and Consurf, all species have evolved only once. This is further illustrated in the Table 12.13. What is evident from the phylogenetic tree branches of the two trees is that yeast UBC1 and its homologues have segregated very well, distinct from all other UBCs.

Table 12.13 Gene duplication in the UBC1 family

Different UBC1s	Number of genes	Different splice /start variants
Homo sapiens	1	
Oryza sativa	1	
Arabidopsis thaliana	1	
Mouse	1	
C.elegans	1	
Plasmodium falciparum	1	
Anopheles gambaie	1	
Neurospora crassa	1	
S.pombe	1	
D. melanogaster	1	
S.cerevisiae	1	
P. yoelli yoelli	1	

Table 12.13 shows that all homologues of yeast UBC1 have appeared only once without any gene duplication.

12.3.12. UBCs and UBE2Vs, TSG101

Figure 12.25.1 MMS2 and UBE2Vs branches generated by Phylip

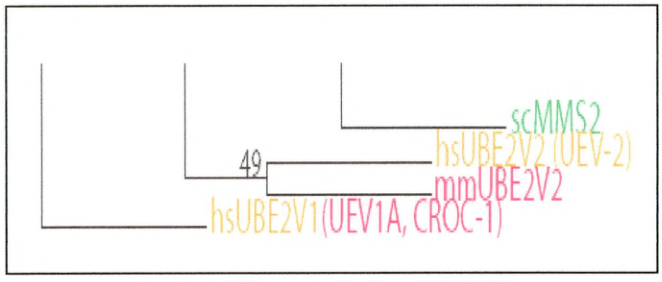


Figure 12.25 shows that the yeast MMS2 and human CROC-1 (UBE2EV1) are very distantly related to all other UBCs as obtained in the Phylip tree.

Figure 12.25.2 TSG101 and UBC9 branch generated by Phylip

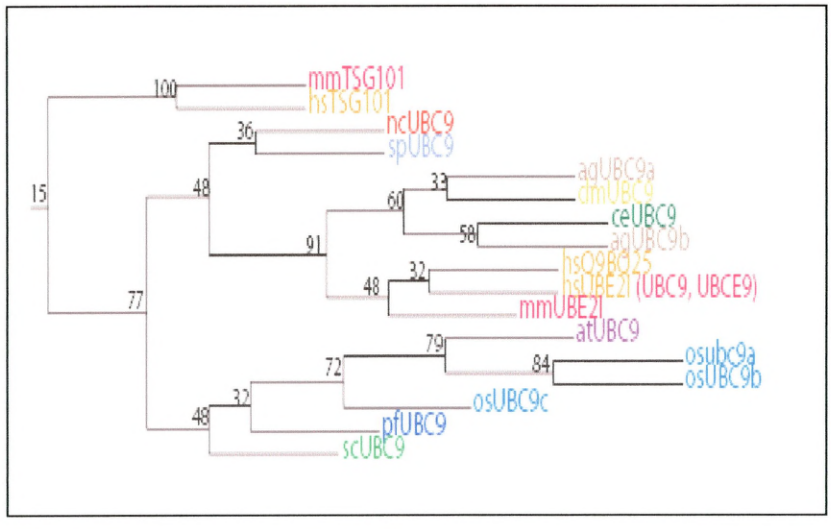


Figure 12.25.2 indicates that TSG101 may be a distant structural relative of the UBC9 family, as represented by the Phylip tree.

Figure12.26 MMS2, TSG101 and UBE2V branch generated by Consurf

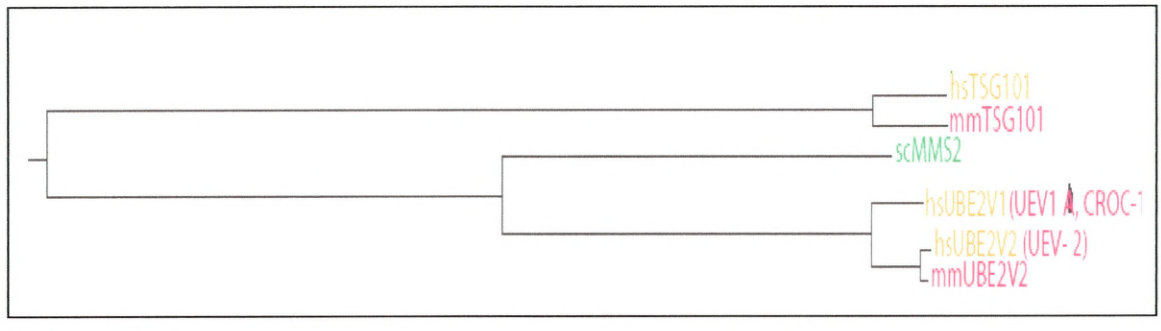


Figure 12.26 indicates that TSG101 and the MMS2/UBE2V family may be more closely related than all other UBCs, as represented by the Consurf tree.

The distant UBC like TSG101 and MMS2/UBE2V families, completely lack a UBC active site i.e. no cysteine or conserved amino acids from either UBC6 or non-UBC6 PROSITE signatures. It is therefore not surprising that Consurf analysis has placed both families on separate adjacent UBC evolutionary sub branches (Figure 12.2). From these Consurf results, this suggests that the TSG101 and the MMS2/UBE2V families had a common ancestor.

Surprisingly the Phylip tree (Figure 12.1) suggests that TSG101 is more closely related to the UBC9 family, than to any member of the MMS2/UBE2V family.

12.3.13. An overall view of the phylogenetic analysis

As can be seen from the constructed phylogenetic trees (Figure 12.1 & 12.2), functional members of the UBC family have been clearly separated into approximately 15 distinct functional branches. As the function of all 13 UBC's in yeast is known, the constructed phylogenetic trees makes it possible to predict the putative function of other plant and animal species that are on the same branch of the tree as their yeast homologue. For example human UBE2J1 and UBE2J2 are on the same phylogenetic branch as yeast UBC6, and which have all been shown to have similar cellular functions (Lenk et al., 2002). Another possible use of this UBC phylogenetic tree is that it can serve to simplify the task of the different eukaryotic nomenclature committees, whereby appropriate names can be ascribed to these branches of related enzymes. This analysis is of particular importance as it has been found that many of the UBC enzymes submitted to the database have either multiple or inappropriate names (see Table 12.14). By using a unified approach based on well constructed phylogenetic trees (Figure 12.1 & 12.2) for all UBC's, it is hoped that the subsequently generated trees will greatly aid scientists in investigating genome function, structure and/or nomenclature in their chosen organism.

12.4. Revised nomenclature of human UBCs

Table 12.14 Table of yeast human UBC nomenclature

YEAST UBCs	Yeast misnomers	Human names and Misnomers
UBC1		**HIP2** (UBE2K) (UBC1_Human)
UBC2	RAD6	**UBE2A** (RAD6A, HHR6A) **UBE2B**(ubc2b, RAD6B, HHR6B)
UBC3	CDC34, DNA6	**CDC34** (UBE2R1) **UBE2R2**
UBC4/5		**UBE2D1** (UbcH5A, UBCH5) **UBE2D2** (Ubch5b) **UBE2D3** (UBCH5C) **UBE2D4** (HBUCE1) **UBE2E1** (Ubch6) **UBE2E2** (UBCH8) **UBE2E3** (UBCE4, UBCH9)
No yeast orthologues		**UBE2L3**(UBCE7, UBCH7) **UBE2L6** (UBCH8)
UBC6	DOA2	**UBE2J1** (NCUBE1, UBC6e, HSPC153, CGI-76) **UBE2J2** (NCUBE2, Ubc6p)
UBC7		**UBE2G1** **UBE2G2**
UBC8		**UBE2H** (UBCH, UBC8)
UBC9		**UBE2I** (UBC9, UBCE9)
UBC10	PEX4, UBCX, PAS2	**UBE2S** (E2EPF)
UBC11		**UBE2C** (UBCH10)
UBC12		**UBE2M** (hUbc12)
UBC13		**UBE2N** (UbcH-ben, MGC8489, BLU)
MMS2		**UBE2V1** (UEV1A , CROC-1) **UBE2V2** (UEV-2, DDVit-1, EDPF-1, MMS2)
No yeast orthologues		**TSG101**

Table 12.14 illustrates all 13 yeast UBCs in the first column, the misnomers of yeast in the second column, and in the third column are all the HGNC nomenclature approved human names of corresponding yeast UBCs in bold, and its misnomers in brackets. All human nomenclature approved names have the prefix UBE2, followed by the alphabets specific to each 13 UBCs.

12.4.1. Yeast UBC1 and HIP2 (UBE2K)

From the phylogenetic trees shown in Figures 12.1 and 12.2, it can clearly be seen that as yeast UBC1 and the human HIP2 (human interacting protein 2) are on the same branch of the UBC tree they are therefore very likely to be orthologues. This was confirmed by using the *Sacchromyces cerevisiae* option in the Ensembl database (www.ensembl.org), which also identifies HIP2 as the predicted orthologue of yeast UBC1. The Human genome organisation (HGNC) nomenclature committee (http://www.gene.ucl.ac.uk/nomenclature/) still uses the name HIP2 and not its previously recommended name UBE2K or UBC1_HUMAN as it is known in UNIPROT (http://www.ebi.uniprot.org/index.shtml). The persistent use of HIP2 (Lee, et al. 2001) is particularly confusing as it bear no sequence or functional resemblance to Huntingtin interacting protein 1 HIP1 (Wanker et al., 1997) and the human intraparietal area 2 (HIP2) (Choi et al., 2006).

12.4.2. Yeast UBC2 and human UBE2A and UBE2B

From Figures 12.1 and 12.2, it can be seen that both human UBE2A and UBE2B are on the same branch as yeast UBC2 and are therefore likely to be orthologues. Again the Ensembl entry for yeast UBC2 agrees with this prediction. Unfortunately the HGNC nomenclature database calls UBE2A and UBE2B as RAD6 and not as a yeast UBC2 orthologues.

12.4.3. Yeast UBC3 (CDC34), human CDC34 (UBE2R1) and UBE2R2

From Figures 12.1 and 12.2, it can be seen that both human CDC34 (UBE2R1) and UBE2R2 are on the same branch as yeast UBC3 and again are likely to be orthologues. Again the Ensembl entry for yeast UBC3 agrees with this prediction.

Like HIP2, HGNC calls human CDC34 as the proper nomenclature name and does not mention its yeast homologue. In addition the HGNC database calls UBE2R1, as an alias of CDC34 which is confusing as the related human protein UBE2R2 is also a functional orthologue of yeast UBC3. An example of the confusion caused by this nomenclature is a recent paper entitled "Ubiquitin-conjugating enzyme 3 delays human lens epithelial cells in metaphase" (Liu et al., 2006), which fails to distinguish which Human orthologue of yeast UBC3, CDC34 (UBE2R1) or UBE2R2, they were investigating.

12.4.4. Yeast UBC4/5, Human UBE2D's and Human UBE2E's

From the phylogenetic trees it can clearly be seen that yeast UBC4, yeast UBC5 and four different human UBE2D proteins (1-4) are on the same branch of the UBC tree

(see Figures 12.1 and 12.2) and are therefore very likely to be orthologues. In support of this assumption the Ensembl database has all four human UBE2D's are quoted as predicted orthologues of both yeast UBC4 and yeast UBC5. Unfortunately the Consurf phylogenetic analyses (Figure 12.2) places two of the UBE2D's (UBE2D1 and UBE2D2) on a different branch from yeast UBC4, yeast UBC5 , human UBE2D3 and human UBE2D4. Consurf phylogenetic analysis is however a less rigorous statistical package designed more to look for structural rather than actual sequence homology. Although the human UBE2E's (1-3) appear on the same evolutionary sub-branch (Figure 12.2) as yeast UBC4, yeast UBC5 they appear not to be as closely related to these yeast proteins as the UBE2D family and are therefore probably not true orthologues of yeast UBC4 and UBC5.

Unfortunately protein biochemists have been especially slow to change over to the (HGNC) UBE2D names and are currently still using the very ambiguous UBCH5a (UBE2D1), UBCH5b (UBE2D2), UBCH5c (UBE2D3) names (Brzovic et al., 2006; Lopez-Avalos et al., 2006; Buchwald et al., 2006). Even more confusingly scientists are still using UBCH6 for UBE2E1 (Buchwald et al., 2006; Karaczyn et al., 2006), UBCH8 for UBE2E2 (Takeuchi, Inoue, and Yokosawa, 2006), and UBCH9 for UBE2E3 (Woods et al., 2004). The use of UBCH8 (UBE2E2) is particularly confusing as the name UBCH8 is currently wrongly being used for the human UBE2L6 protein (Tripathi, and Chaudhuri, 2005)!

12.4.5. Human UBE2L3 and UBE2L6

From the phylogenetic trees shown in Figures 12.1 and 12.2, it can clearly be seen that the related human UBE2L3 (UBCH7) and UBE2L6 (UBCH8) proteins are on their own branch of both the Phylip and the Consurf trees. There appears to be no known yeast orthologues to these proteins. UBE2L6 has been shown to be involved in the ubiquination of the cell cycle control protein P53 (Huang et al., 2006). Again many protein biochemists are still currently using the highly confusing names UBCH7 and UBCH8 for UBE2L3 and UBE2L6 respectively (Fortier and Kornbluth, 2006; Takeuchi, Inoue and Yokosawa, 2006; Wong et al., 2006; Takeuchi et al., 2005).

12.4.6. Yeast UBC7 and human UBE2G

From the phylogenetic trees shown in Figures 12.1 and 12.2, it can clearly be seen that as yeast UBC7 and the two human UBE2G's (1 and 2) are on the same branch of the

UBC tree they are therefore very likely to be orthologues. This was confirmed by using the *Sacchromyces cervisiae* option in the Ensembl database (www.ensembl.org). Unfortunately many biochemists confusingly still insist on using the yeast UBC7 name to describe UBE2G2, clearly ignoring the fact that another human orthologue, UBE2G1, also exists for yeast UBC7 (e.g. Briggman et al., 2005).

12.4.7. Yeast UBC8 and human UBE2H

The phylogenetic trees shown in Figure 12.1 and 12.2, clear identifies yeast UBC8 and the human UBE2H in the same phylogenetic sub branch, which signifies that they are most likely to be orthologues, which is further confirmed by the Ensembl database (www.ensembl.org).

12.4.8. Yeast UBC9 and UBE2I

The phylogenetic trees shown in Figures 12.1 and 12.2, identifies yeast UBC9 and the human UBE2I in the same phylogenetic sub branch. This signifies that UBE2I is most likely to be an orthologue of yeast UBC9, which was further confirmed by the Ensemble database (www.ensembl.org).

12.4.9. Yeast UBC10 and human UBE2S

The yeast UBC10 (Alias UBCX, Pex 4, and Pas2) and the human UBE2S appears in the same phylogenetic sub branch as in Figures 12.1 and 12.2, which specifies that it is most likely an orthologue. Although the human UBE2S name is now widely used and accepted unfortunately yeast UBC10 is also known as several other names e.g. UBCX, Pex4 and Pas2. Disappointingly the Ensembl database uses Pex4 as the recognised name for yeast UBC10 and has not yet named human UBE2S as its orthologue. The persistent use of the name pex4 is not logical and can be confusing e.g. Eckert and Johnsson's (2003) paper is entitled "Pex10p links the ubiquitin conjugating enzyme Pex4p to the protein import machinery of the peroxisome." this could be written more simply and accurately as "Pex10p links UBC10 to the protein import machinery of the peroxisome."!

12.4.10. Yeast UBC11 and human UBE2C

From the phylogenetic tree it could be said that human UBE2C is probably a yeast UBC11 orthologue as they appear in the same phylogenetic sub branch. This is further confirmed by the Ensembl database (www.ensembl.org), which quotes human UBE2C as a predicted orthologue for yeast UBC11. Unfortunately the, old and confusing, alias

name UBCH10, and not human UBE2C, is currently still being used by biochemists (e.g. Fortier and Kornbluth, 2006).

12.4.11. Yeast UBC12 and human UBE2M

Once again both the yeast UBC12 and the human UBE2M are in the same phylogenetic sub branch as in Figures 12.1 and 12.2, and is further confirmed by the Ensemble database (www.ensembl.org). From this it could be said that UBE2M is most likely an orthologue of yeast UBC12.

12.4.12. Yeast UBC13 and human UBE2N

Yeast UBC13 and the human UBE2N are on the same phylogenetic sub branch shown in Figures 12.1 and 12.2, and which is further confirmed by the Ensemble database (www.ensembl.org). Hence it is most likely that UBE2N is an orthologue of yeast UBC13. UBE2N has its aliases mentioned in the HGNC database as (UbcH-ben, MGC8489, BLU).

12.4.13. Yeast MMS2 and the human UBE2V, and the TSG101

The yeast MMS2 and the human UBE2Vs (UBE2V1 and UBE2V2) are on the adjacent sub branches of the phylogenetic tree. The aliases of the human variants of yeast MMS2 (UBE2V1 and UBE2V2) are UBE2V1 (Uev1A, CROC-1) (Andersen et al., 2005), and UBE2V2 (UEV-2, DDVit-1, EDPF-1, MMS2) (Lewis et al., 2006). The TSG101 appears in the sub branch on its own in the phylogenetic tree as in Figures 12.1 and 12.2.

12.4.14. Yeast UBC6 and the human UBE2J

My supervisor Dr. Lester previously identified two potential human orthologues of yeast UBC6 (Lester et al., 2000), which he and his co-workers named NCUBE1 (now UBE2J1) and NCUBE2 (now named UBE2J2). Weissman et al. (2001) later identified a mouse orthologue of yeast UBC6, which he named MmUBC6 (now named UBE2J2). The next year Lenk et al. (2002) proved experimentally that two human proteins UBC6 (now named UBE2J2) and UBC6e (now named UBE2J1) were the true functional orthologues of yeast UBC6. To try and clear up the possible future confusion over the 2 human yeast UBC6 orthologues my supervisor and I contacted Hester Wain of the HGNC, who in 2003, deemed that NCUBE was not a suitable nomenclature name, but

offered us an option of several UBE2 names. As the letter J was still available for UBCs, and being the initial of my first name Joy, my supervisor and I chose UBE2J1 for NCUBE1 and UBE2J2 for NCUBE2. Unfortunately many authors still insist on using the old, redundant and confusing nomenclature (e.g. Oh et al., 2006).

12.4.15. Conclusions from HGNC approved UBC nomenclature

As can be seen from the above section there are still many anomalies for both the yeast and human nomenclature committees to resolve. Where there is possible ambiguity it would be preferable if authors used both their preferred nickname and in brackets the approved HGNC nomenclature e.g. HIP2 (UBE2K) and CDC34 (UBE2R1). These examples immediately show that as well as being either a hutington interacting protein (HIP2) or a cell cycle protein (CDC34), these proteins are also ubiquitin conjugating enzymes! Moreover the nomenclature UBE2R1 suggests that another closely related paralogue exists, namely UBE2R2, which is likely to be also a cell cycle control protein.

12.5 Structural Conservation results from Consurf analysis

The results of the UBE2J2 Consurf analysis are given in three sections. Firstly the Consurf phylogenetic tree (Figure 12.2), secondly the three dimensional structure of the protein UBE2J2 (PDB ID- 2F4W), with its conservation parameters in respective colours (Figure 12.27A and Figure 12.27B) and finally the same conservation of the protein UBE2J2, shown on the peptide sequence of UBE2J2 (Figure 12.27C). The significance of the conservation results from the Consurf analysis (Figure 12.27C) is that it mostly correlates with the highly conserved amino acids in our proposed UBC6 PROSITE signature (Figure 12.27D).

As was said previously, Consurf could be used to identify conserved surface features of homologous proteins. Figure 12.27 was generated by using the human homologue of yeast UBC6 (UBE2J2), as a template onto which all other UBC structural conservations were projected. In this way conserved structural features among all UBC homologues was identified.

Figure 12.27 Evolutionary conservation on the 3D structure the UBE2J2 protein (2F4W)

Figure 12.27A **Figure 12.27B**

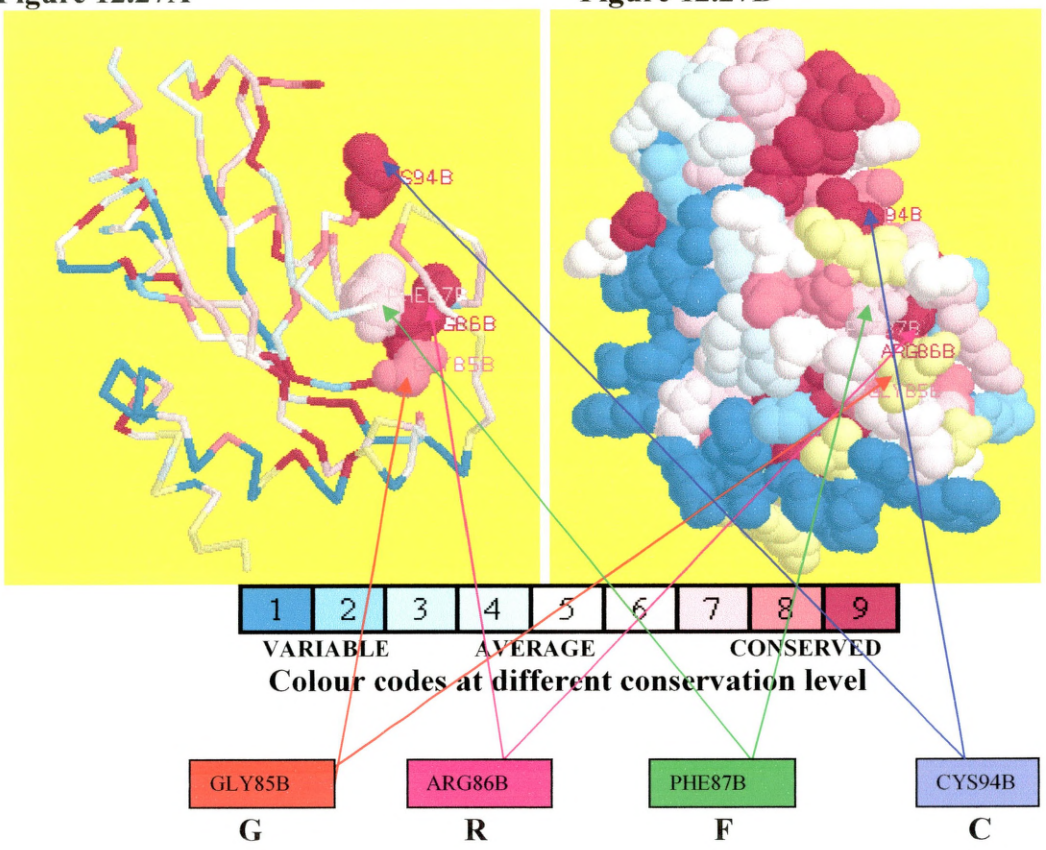


Figure 12.27C

Colour codes at different conservation level

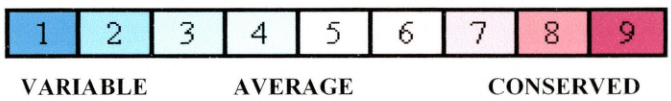


Figure 12.27D Pairwise alignment of UBC6 (UBE2J2) with a non UBC6 [UBC9]

Our proposed PROSITE signature of UBC6:

T-[PAR]-[NS]-G-R-F-x(3)-[KTE]-[RK]-[LIV]-C*-[LMS]-[ST]-[IMF]-[ST]-x(2)-H-[PK]

9 2 9 8 9 7 -5 8 5 8 9 8 **These are the conservation scores of corresponding amino acids in the PROSITE signature**

"score 5" is not coloured white as otherwise it would not be visible

```
hs2J2_V1: HYVVRGPEMTPYEGGYYHGKLIFPREFPFKPPSIY----MITPNGRFKCN
          ...:.|.:.||:|||.:..:::|..::|..||...    :..||| ....
hsUBE2I : ecaipgkkgtpwegglfklrmlfkddypssppkckfepplfhpng-vpfg

hs2J2_V1 :TRLCLSITDFHPDTWNPAWSVSTILTGLLSFMVEKGPTLGSIETSDFTKR
          | :||||.:...| |.||.::..||.|:...:.|  |.:
hsUBE2I : t-vclsileedkd-wrpaitikqillgiqellne--pni----------
```

The PROSITE signature of all other UBCs [e.g.UBC9 (UBE2I)] except UBC6--

[FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x- [LIV]-C*-[LIV]-x-[LIV]

Figures 12.27 (A-D), shows the various levels of conservation of the amino acid sequence mapped on the 3D structure of the protein UBE2J2. Figure 12.27A and 12.27B are the same, as they are the conservation of amino acids onto the 3D structure of UBE2J2. Figure 12.27A shows the overall conservation of UBCs onto the 3D structure of UBE2J2 with the active site residue Cysteine, and 3 other amino acids of its PROSITE signature region highlighted. Figure 12.27B also shows the overall conservation of amino acids, but with a better view of conservation. Figure 12.27C shows the same conservation of amino acids at the sequence level. There is a gap at the region of the PROSITE signature of the peptide sequence, just after the active site residue Cysteine. This is because, although those amino acids exist, their coordinates could not be obtained from the X-ray diffraction of the crystal structure of 2F4W. Also represented is the index of the different colour codes of conservation at various conservation levels from 1 to 9 that are projected onto the 3D structure. Figure 12.27D shows a pairwise alignment of UBC (UBE2J2) and a non-UBC6 [UBC9 (UBE2I)], illustrating the differences in their PROSITE signatures.

From Figures 12.27 (A, B, C), it can clearly be seen that the invariant UBC active site cysteine, that is present in both the non-UBC6 and our UBC6 PROSITE signature is not surprisingly in the highest conserved category (coloured maroon, conservation 9) on both the 3D structural diagrams and the 2D Consurf results. Surprisingly the GRF motif, which is only found in UBC6 like enzymes, and not non-UBC6s, appears also to be in the highly conserved categories (between conservation 7-9). This result appears to indicate that although the amino acids around the cysteine active site are very different between the UBC6 and non-UBC6 families, the amino acid properties and therefore possibly the structure of the two families active sites, have remained similar.

In support of this hypothesis our UBC6 PROSITE signature states that after the conserved amino acids G, R, F, there can be any three amino acids (e.g. KCN in UBE2J2). Interestingly these amino acids are grouped in less conserved categories (conservation 3-5). This result is consistent with these amino acids being less important for determining the structure and function of the UBC6 family active site.

The adjacent highly conserved [KTE] of the UBC6 PROSITE signature, again appears to comply with the Consurf results, as the T (tyrosine) in UBE2J2 is also in the highly conserved category 8.

From Figure 12.27C it can therefore be seen that generally the more conserved an amino acid is (e.g. between 1 and 3 options), in our UBC6 PROSITE signature it is also likely to be in the highly conserved category following our Consurf analysis using all UBC's. There are however two notable exceptions. Firstly in UBE2J2 the proline residue, which is part of the [PAR] part of the UBC6 PROSITE signature, has a very low conservation 2 from the Consurf analysis. This would suggest that the structure of the active site of UBE2J2 is therefore likely to differ considerably from other UBCs, due to this proline residue, which is known to introduce distortions in structures. The second notable exception is the highly conserved positively charged amino acids arginine that forms part of the UBC6 PROSITE signature [RK]. These positively charged amino acids are completely absent from non UBC6 UBCs PROSITE signature. A low conservation score of 5 is therefore found for this arginine residue in UBE2J2. The presence of a positively charged amino acid near the active site of (UBE2J2) UBC6 like enzymes, that is absent from other UBCs, may also produce a difference in its structure and therefore in their substrate specificity.

12.6 Conclusion

To conclude it can clearly be seen from the Phylip 3.6 generated phylogenetic tree that the 13 yeast UBCs and the UBC-like protein yeast MMS2, with their homologues have been resolved into 15 individual orthologous branches. Two of the 15 branches contain no yeast orthologues. One of these branches contains human UBE2L3 and UBE2L6, while the other branch contains the UBC like protein TSG101 that lacks a cysteine active site. From the distinct divergence of the 15 branches it is likely that all the UBC and UBC like proteins have evolved from a common ancestor, by gene duplication. Following gene duplication these genes have mutated and have evolved different functions. These results indicate that any new putative UBCs can be assigned to the correct orthologous group by using these phylogenetic approaches. This approach should be particularly useful in assigning correct nomenclature in plant species that have undergone considerable gene duplication of most ancesteral UBCs (see gene duplication tables in previous sections) e.g. yeast UBC7 appears to have 5 different orthologues in *Arabidopsis thaliana* (Table 12.6).

Finally these phylogenetic trees should help in the revision of the nomenclature for non-yeast UBCs.

12.7 Future prospects

The phylogenetic tree developed in this study can now be taken as a reference tree, acting as a backbone for any future analysis. With the increasing number of eukaryotic genomes being sequenced many new UBC and UBC-like peptides of unknown function will be identified, so by adding these newly identified sequences to the sequences used in this phylogenetic analysis and carrying out a new phylogenetic analysis, the new UBC sequences should appear in branches, where the functional and evolutionary relationships can be deduced. This analysis should also help in assigning the correct and less confusing names to new UBC peptides.

CHAPTER THIRTEEN

RESULTS & DISCUSSION OF PHYLOGENETIC ANALYSIS OF $TAF_{II}250$

13.1. Results and discussion of phylogenetic analysis of $TAF_{II}250$

Pham and Sauer (2000) previously showed that the *Drosophila* transcription factor $TAF_{II}250$ possesses both E1 Ubiquitin-activating and E2 ubiquitin conjugating activity and is able to mono-ubiquinate histone H1 in chromatin. From this paper the amino acid positions on the $TAF_{II}250$ protein for the E1 active site were demonstrated experimentally using previously identified *Drosophila* mutants (Pham and Sauer, 2000; Wassarman et al., 2000). The exact position of the E2 active site on $TAF_{II}250$ was not, however, elucidated from these experiments, but was shown to lie between amino acids 612 and 1140. Therefore in order to try and identify the exact location of the UBC (E2) active sites in the *Drosophila* $TAF_{II}250$ protein a search was carried out for a UBC PROSITE signature in this and other regions of the protein (Pham and Sauer, 2000).

All UBC's contain a cysteine residue in their active site, which binds ubiquitin before it is added to a specific protein substrate (Passmore and Barford, 2004; Lester et al., 2000). In addition, with the notable exception of yeast UBC6 and its homologues (Lester et al., 2000), most UBCs have the PROSITE signature [FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-**C**-[LIV]-x-[LIV] at their active sites. Any putative *Drosophila* $TAF_{II}250$ UBC active site is therefore likely to also contain both a cysteine residue and similar surrounding amino acid residues to known UBC's. For example most UBCs have, an invariant Glycine, 2 amino acids upstream to the active site cysteine residue.

Of the 2065 amino acids in *Drosophila* $TAF_{II}250$ only 20 were found to be cysteine residues. Of these 20 cysteine residues only one cysteine, at position 742, has a Glycine residue in close proximity, two amino acids upstream (at position 740). Using both protein sequences from $TAF_{II}250$ and selected UBCs [including CROC1 (UBE2V1)], from different species, clustalW was used and manually aligned to create a multiple sequence alignment (see Figures 13.1.1 & 13.1.2), in order to see if this putative UBC active site was conserved in other metazoan orthologues of *Drosophila* TAF250.

As can be seen from Figure 13.1.1, approximately 70% of the conserved UBC PROSITE amino acids (indicated by an asterix) are also conserved in the putative TAF250 UBC active sites for both *Drosophila* and *Apis mellifera*. It is highly likely therefore that the area surrounding and including cysteine 742 in drosophila and cysteine 631 in honeybee $TAF_{II}250$ are the functional UBC active sites, in these species. Partial $TAF_{II}250$ sequences were obtained for the Red flour beetle and silkworm, which

also showed similar homology. Surprisingly the essential cysteine in the putative E2 active site for $TAF_{II}250$ in these insect species appeared to be absent in all other known $TAF_{II}250$ orthologues including human and mosquito. This suggested that UBC activity is probably absent from $TAF_{II}250$ in most species. Interestingly the sequence surrounding and including the only cysteine located between 1080 and 1137, where Pham and Sauer (2000) previously predicted a UBC active site, in *Drosophila* $TAF_{II}250$, showed no homology to the UBC PROSITE signature.

Using the multiple sequence alignment shown in Figure 13.3, a phylogenetic tree was constructed, using Phylip 3.6 (see figure 13.2). From this tree it can be hypothesised that both UBCs and UBC like enzymes such as CROC1 (UBE2V1) shared a common structural ancestor with the $TAF_{II}250$ family. In support of this hypothesis it has been previously shown that CROC1 (UBE2V1), which lacks a UBC active site, possesses both transcriptional activity and a homologous $TAF_{II}250$ like domain (Rothofsky and Lin 1997), (see Figure 13.1.1). Assuming that $TAF_{II}250$ and the UBC family do indeed share a common structural ancestor, convergent evolution probably occurred later when the UBC family of enzymes and the $TAF_{II}250$ protein in several different insect species, independently evolved a functional UBC active site. This UBC active site in $TAF_{II}250$ appears not to have evolved in mosquitoes, vertebrates or plants (data not shown). Mono-ubiquitination of histone H1 in most eukaryotic species is therefore probably carried out by other UBC (E2) and possibly E3 ubiquitin ligases. For example it has recently been shown that human testis Histones including H1, have been shown to be ubiquitinated by the UBCs: UBE2D2, UBE2L3 and an ubiquitin ligase, $E3^{Histone}$ (Liu, Oughtred and Wing, 2005; Hass, Bright and Jackson, 1988).

Figure 13.1.1

$TAF_{II}250$ & UBC block

```
B.taurus          :ENGPLMMQVG-MATKIK
P.cynocephalus    :ENGPLMMQVG-MATKIK
M.fascicularis    :ENGPLMMQVG-MATKIK
M.musculus        :ENGPLMMQVG-MATKIK
R.norvegicus      :ENGPLMMQVG-MATKIK
M.auratus         :ENGPLMMQVG-MATKIK
C.familiaris      :ENGPLMMQVG-MATKIK
H.sapiens         :ENGPLMMQVG-MATKIK
H.lar             :ENAPLMMQIG-MATKIN
P.pygmaeus        :ENGPLMMQVG-MATKIK
P.troglodytes     :ENGPLMMQVG-MATKIK
G.gorilla         :ENGPLMMQVG-MATKIK
T.nigroviridis    :EYPPLFMQVG-MASKIK
G.gallus          :ENAPLMMQVG-MATKIK
D.rerio           :EYPPLHMQVG-MASKIK
hsCROC1           :NMNGVNSSNGVVDPRAI
c.elegans         :EHPVILSQPG-MASKMK
C.briggsae        :EHPVLLGQPG-MASKIK
A.gambiae         :EHPPLMNQVG-MASKIK

D.melanogaster    :EHPPLINQVG-MCSKIK
D.pseudoobscura   :EHPPLMNQVG-MCSKIK
A.mellifera       :EHPPLMNQVG-MCSKVK
scUBC4            :YHPNIN-ANGNICLDIL
scUBC13           :YHPNID-RLGRICLDVL
scUBC1            :YHPNISSVTGAICLDIL
atUBC1            :FHPNIY-ADGSICLDIL
scUBC2RAD6        :FHPNVY-ANGEICLDIL
hsUBC9            :FHPNVY-PSGTVCLSIL
hsUBC12           :YHPNID-LEGNVCLNIL
                    ** ******* * **
```

Figure 13.1.2

$TAF_{II}250$ & CROC1(UBE2V1) block

```
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRNQEKEKLRGPPE
RRIQEQLRRLKRNQEKEKLRGPPE
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRSQEKEKLKGPPE
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRNQEKDKIKGPPE
RRIQEQLRRLKRNQEKEKLKGPPE
RRIQEQLRRLKRNQEKDRFKGPPE
KVVLQELRRLMMSKENMKLPQPPE
RRLQDQIRRMKKNEEKAAHKVQKM
RRLQDQARRMKKNEEKQQAKTPKQ
RRIQEQLRRIKRNQQK--IGMQMQ

RRIQEQLRRIKRNQERERLAQLAQ
RRIQEQLRRIKRNQERERLAQLAQ
RRIQEQLRRIKRNQERERMLGGPM
------------------------
------------------------
------------------------
------------------------
------------------------
------------------------
------------------------
```

```
[FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x-[LIV]
```

The PROSITE signature of UBCs, cysteine **(C)** being the ubiquitin active site residue.

Figure13.1.1 Shows a *clustalW* multiple sequence alignment of the putative UBC active site block in $TAF_{II}250$ and selected UBCs [including the UBC like enzyme CROC1(UBE2V1)]. It can be seen that there is a conservation of approximately 70% of the UBC PROSITE signature (shown by *), including the cysteine active site in the drosophila and honeybee $TAF_{II}250$ in this region. The cysteine active site is however, absent in all other species in this alignment. Figure 13.1.2 shows the conservation of the $TAF_{II}250$ like motif in the human CROC1 enzyme compared to the $TAF_{II}250$ protein sequences of several different animal species.

Figure 13.2 Unrooted phylogenetic tree generated from the MSA of $TAF_{II}250$, UBC & CROC1 (UBE2V1)

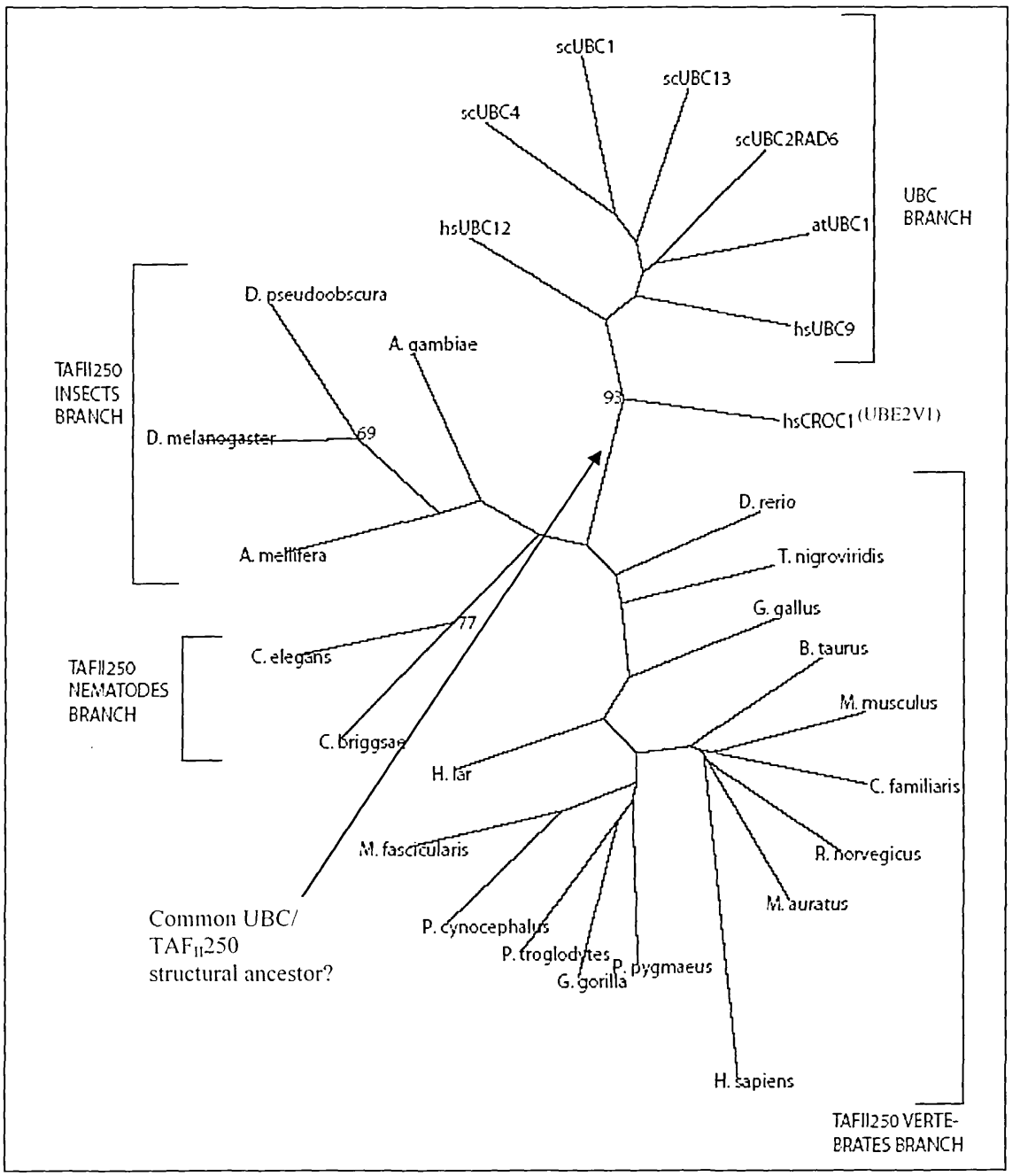


Figure 13.2 Shows a phylogenetic tree based on the larger clustalW generated multiple sequence alignment (see Figure 13.3 on the next page for the whole alignment), using Phylip3.6 neighbour joining method and 100 bootstrap replicates.

Figure 13.3 **The whole MSA of $TAF_{II}250$, UBC & CROC1 (UBE2V1)**

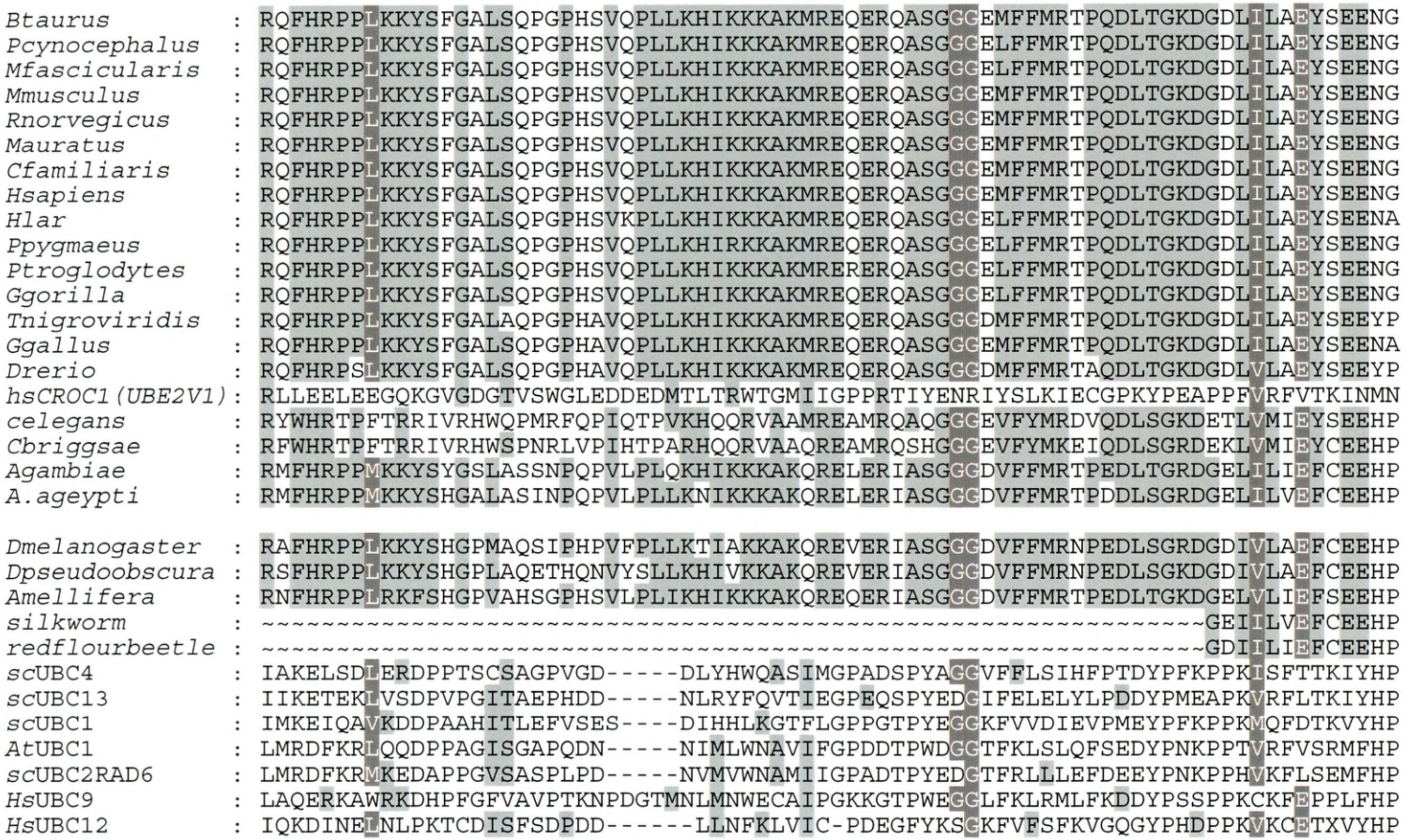

Btaurus : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMREQERQASGGEMFFMRTPQDLTGKDGDLILAEYSEENG
Pcynocephalus : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMREQERQASGGELFFMRTPQDLTGKDGDLILAEYSEENG
Mfascicularis : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMREQERQASGGELFFMRTPQDLTGKDGDLILAEYSEENG
Mmusculus : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMREQERQASGGEMFFMRTPQDLTGKDGDLILAEYSEENG
Rnorvegicus : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMREQERQASGGEMFFMRTPQDLTGKDGDLILAEYSEENG
Mauratus : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMREQERQASGGEMFFMRTPQDLTGKDGDLILAEYSEENG
Cfamiliaris : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMREQERQASGGEMFFMRTPQDLTGKDGDLILAEYSEENG
Hsapiens : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMREQERQASGGEMFFMRTPQDLTGKDGDLILAEYSEENG
Hlar : RQFHRPPLKKYSFGALSQPGPHSVKPLLKHIKKKAKMREQERQASGGELFFMRTPQDLTGKDGDLILAEYSEENA
Ppygmaeus : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIRKKAKMREQERQASGGELFFMRTPQDLTGKDGDLILAEYSEENG
Ptroglodytes : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMRERERQASGGELFFMRTPQDLTGKDGDLILAEYSEENG
Ggorilla : RQFHRPPLKKYSFGALSQPGPHSVQPLLKHIKKKAKMREQERQASGGELFFMRTPQDLTGKDGDLILAEYSEENG
Tnigroviridis : RQFHRPPLKKYSFGALAQPGPHAVQPLLKHIKKKAKMREQERQASGGDMFFMRTPQDLTGKDGDLILAEYSEEYP
Ggallus : RQFHRPPLKKYSFGALSQPGPHAVQPLLKHIKKKAKMREQERQASGGEMFFMRTPQDLTGKDGDLILAEYSEENA
Drerio : RQFHRPSLKKYSFGALSQPGPHAVQPLLKHIKKKAKMREQERQASGGDMFFMRTAQDLTGKDGDLVLAEYSEEYP
hsCROC1(UBE2V1): RLLEELEEGQKGVGDGTVSWGLEDDEDMTLTRWTGMIIGPPRTIYENRIYSLKIECGPKYPEAPPFVRFVTKINMN
celegans : RYWHRTPFTRRIVRHWQPMRFQPIQTPVKHQQRVAAMREAMRQAQGGEVFYMRDVQDLSGKDETLVMIEYSEEHP
Cbriggsae : RFWHRTPFTRRIVRHWSPNRLVPIHTPARHQQRVAAQREAMRQSHGGEVFYMKEIQDLSGRDEKLVMIEYCEEHP
Agambiae : RMFHRPPMKKYSYGSLASSNPQPVLPLQKHIKKKAKQRELERIASGGDVFFMRTPEDLTGRDGELILIEFCEEHP
A.ageypti : RMFHRPPMKKYSHGALASINPQPVLPLLKNIKKKAKQRELERIASGGDVFFMRTPDDLSGRDGELILVEFCEEHP
Dmelanogaster : RAFHRPPLKKYSHGPMAQSIPHPVFPLLKTIAKKAKQREVERIASGGDVFFMRNPEDLSGRDGDIVLAEFCEEHP
Dpseudoobscura : RSFHRPPLKKYSHGPLAQETHQNVYSLLKHIVKKAKQREVERIASGGDVFFMRNPEDLSGKDGDIVLAEFCEEHP
Amellifera : RNFHRPPLRKFSHGPVAHSGPHSVLPLIKHIKKKAKQREQERIASGGDVFFMRTPEDLTGKDGELVLIEFSEEHP
silkworm : ~~~

Figure 13.3 The whole MSA of $TAF_{II}250$, UBC & CROC1(UBE2V1) continued:

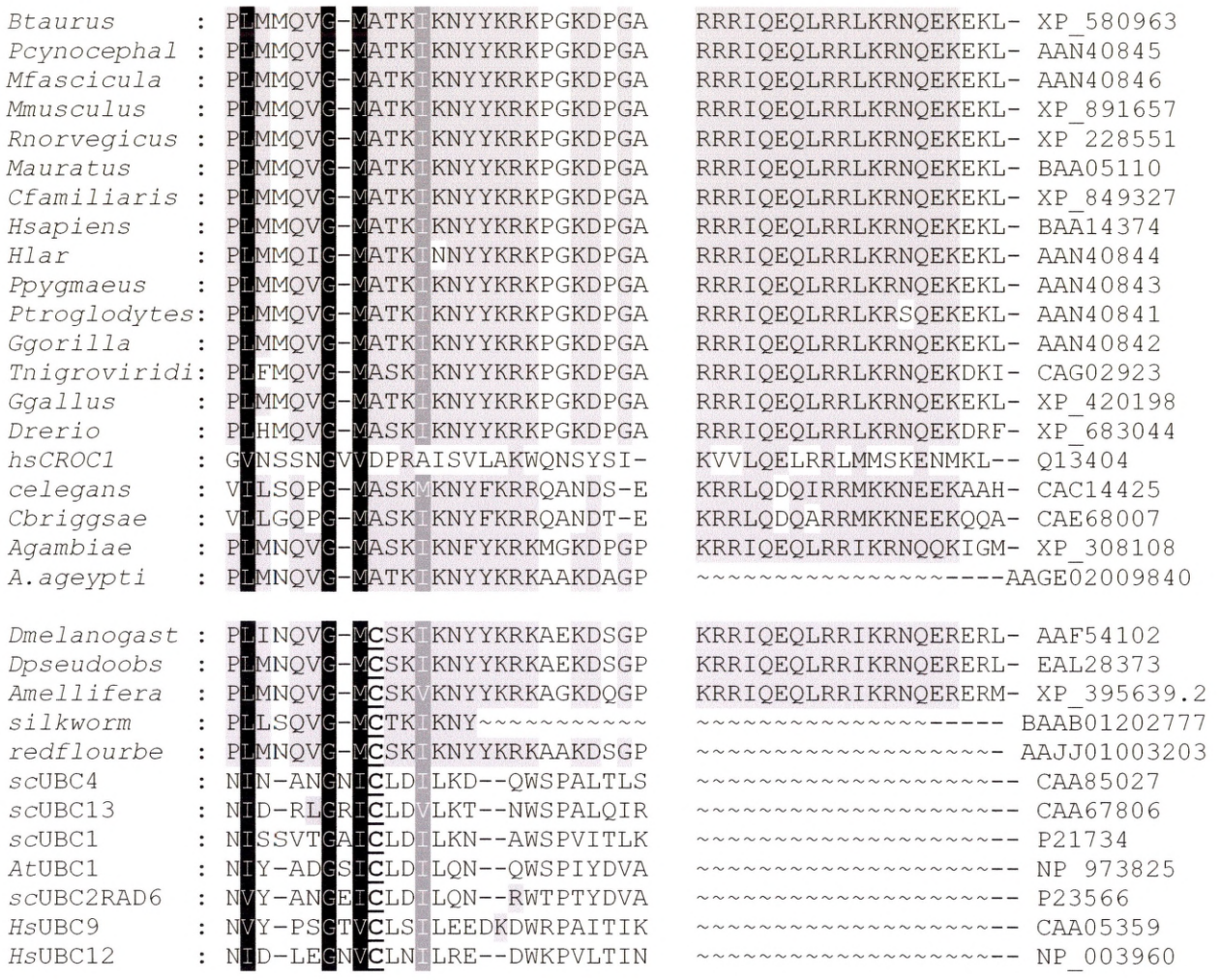


Figure 13.3 is a clustalW (http://www.ebi.ac.uk/clustalw/index.html) and manually aligned (using GENDOC http://gendiapo.sourceforge.net/) multiple sequence alignment of selected $TAF_{II}250$, UBC and UBC like proteins CROC1 (UBE2V1). The Cysteine in the UBC active site is underlined (C). The red flour beetle and silkworm $TAF_{II}250$ were excluded from the phylogenetic analysis as only partial predicted genomic translations were available.

13.2. Conclusion

Previous laboratory experiments have shown that the *Drosophila*, transcription factor, $TAF_{II}250$ has ubiquitin conjugating (UBC) enzyme activity, however, these experiments, failed to identify the exact location of this UBC active site. A putative UBC active site was identified, in the *Drosophila* $TAF_{II}250$ amino acid sequence. Interestingly this UBC active site was found to be absent from most other species studied, with the notable exceptions of *Drosophila melanogaster* and *Apis mellifera*. These results suggest that the UBC active site in these species is likely to have occurred due to convergent, rather than divergent evolution.

CHAPTER FORTEEN

RESULTS AND DISCUSSION OF STRUCTURE DETERMINATION BY X-RAY CRYSTALLOGRAPHY AND CIRCULAR DICHROISM SPECTROSCOPIC ANALYSIS

14.1. Results of the expression of UBE2J1 for X-ray crystallography

14.1.1. PCR of UBE2J1 fragment

As described in the materials and methods section (chapter 9), a PCR reaction of the UBE2J1 fragment was carried out with the normal brain cDNA, and ligated into an expression vector. This plasmid was then successfully transformed into *E. coli*. The expressed protein was then purified and attempts were then made to crystallise this protein fragment.

Figure 14.1 Gel electrophoresis of PCR reaction products and the plasmids

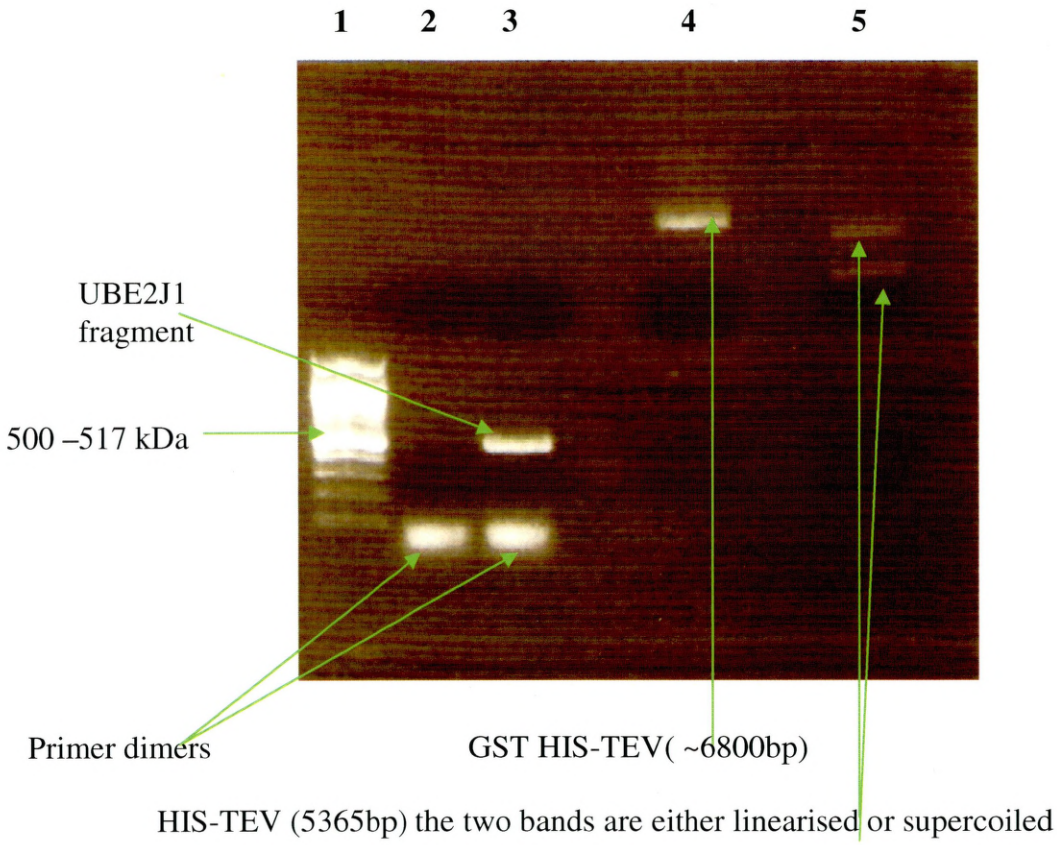


Figure 14.1 is a gel electrophoresis of the PCR reaction. Lane 1 on the gel diagram is the 100bp DNA ladder, lane 2 is the blank control, lane 3 is the PCR reaction product of the UBE2J1 fragment. In lane 4 is the GST HIS-TEV30A plasmid, which is shown in the materials and methods. In lane 5 is the HIS-TEV30A plasmid.

14.1.2. Digestion by the restriction enzymes

The PCR reaction product of the UBE2J1 fragment, and also the two plasmids as mentioned previously, were then subjected to digestion before ligation into the plasmid vector, by the restriction enzymes *BamH1* and *EcoR1*, procedure described in the materials and methods.

The digested products were later run on an agarose gel as shown in Figure 14.2.

Figure 14.2 Gel electrophoresis of the digested products

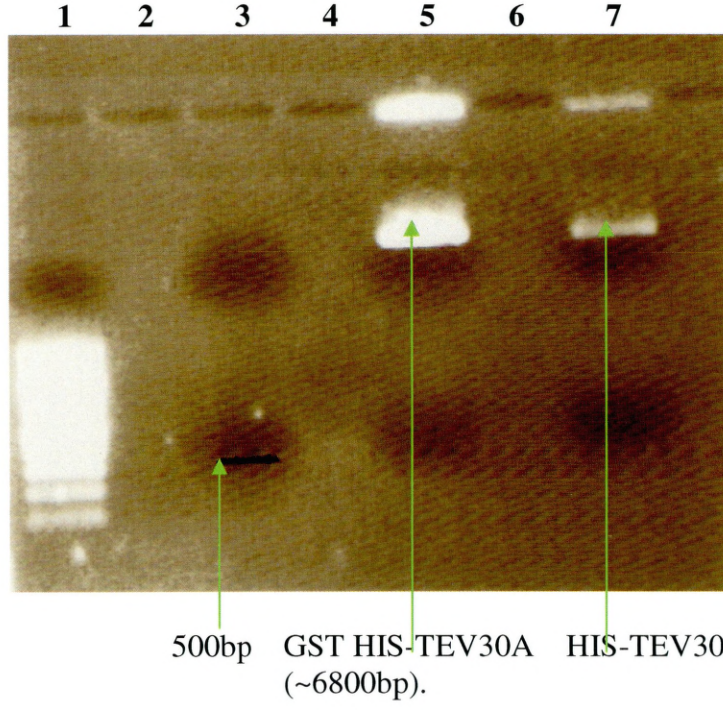


Figure 14.2 is the gel elecrophoresis of the digested products in different lanes. In lane 1 is the 100bp DNA ladder, lane 3 is the BamH1+ EcoR1 digest of the PCR product, lane 5 is the BamH1 + EcoR1 digest of GST HIS-TEV30A plasmid, and lane 7 is the EcoR1 + BamH1 digest of HIS-TEV30A plasmid. The gel was made of agarose (1%) with 10 µl of 0.0005% w/v ethidium bromide and *was run at 40-60V for 1 hour.*

14.1.3. Digestion by the restriction enzymes of the digested UBE2J1 fragment and HISTEV30A plasmid

Following ligation of UBE2J1 fragment into HISTEV30A and GST HISTEV30A plasmids, a *BamH1* and *EcoR1* digestion was carried out to prove that the UBE2J1 PCR fragment had been successfully ligated. As can been seen from Figure 14.3, that the UBE2J1 fragment was present in several HISTEV30A samples, unfortunately no ligation of the UBE2J1 fragment was achieved with the GST HISTEV plasmid (data not shown).

Figure 14.3 Gel electrophoresis of the digestion by the restriction enzymes

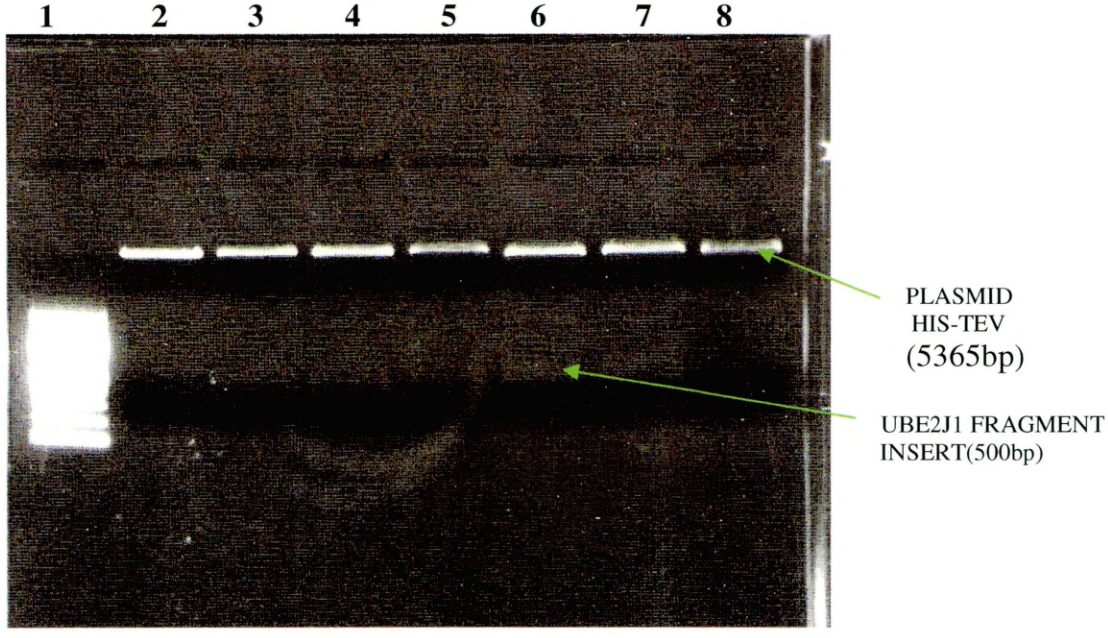


Figure 14.3 illustrates the appearance of the two bands, the plasmid vector (HIS-TEV) and the UBE2J1 fragment in each of the lanes, clearly stating that the desired ligation has occurred, and has not self ligated. Lane 1 is the marker and lanes 3 to 7 are the ligations which have been digested by the restriction enzymes where it clear shows the plasmid HIS-TEV and the UBE2J1 fragment.

Table 14.1.1 Quantity of standard BSA and its O.D. at 595nm

Amount of BSA taken	Absorbance at 595
2μg	0.135
5μg	0.301
10μg	0.536

Table 14.1.1 shows the quantities of standard (BSA) sample taken and its absorbance at 595 nm for determining the standard curve.

Table 14.1.2 Quantity of test sample taken and its O.D. at 595nm

Quantity of test sample taken	Absorbance of test sample	Concentration Of the protein
2μl	0.703	6.6 mg/ml
5μl	1.326	-
10μl	1.876	-

Table 14.1.2 shows the quantities of the protein sample taken and its absorbance at 595 nm, for the determination of its concentration. From the standard curve the concentration of protein sample was obtained by extrapolating on the graph the absorbance of protein sample to its corresponding concentration. Here 0.703 was considered as the absorbance is less than 1 and its concentration was estimated from the standard curve.

From the standard graph plotted, the corresponding concentration of the protein sample from its absorbance of 0.703 is 13.2μg/2μl. So the final concentration is 6.6 mg/ml.

Theoretically 1μg of TEV protease cleaves 1 mg of protein.
Hence various concentrations of TEV protease was mixed with the protein samples and incubated at room temperature for 4 -5 hours and then run on the gel to estimate which concentration of TEV protease cleaved the best. The result is illustrated in Figure 14.4.

Figure 14.4 **TEV-protease clevage trials**

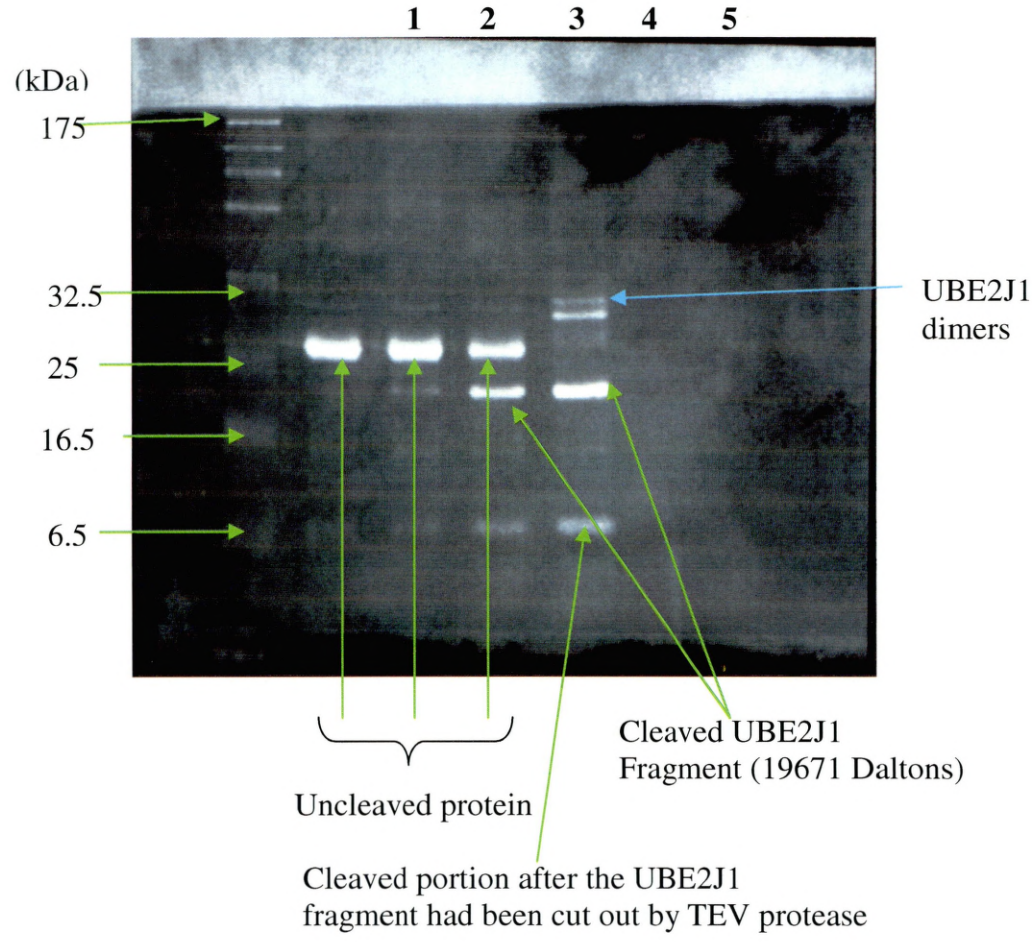


Figure 14.4 illustrates TEV protease cleavage at various concentration. Lane 1 is the protein marker, lane 2 is the uncleaved protein which is a control, lanes 3, 4, 5 are the TEV protease cleave of the protein in concentrations of 10 ng/µl, 100ng/µl, and 1µg/µl. It was found that the lane 5 of the TEV protease concentration of 1µg/µl is the best concentration for complete cleavage of the Histidine tagged protein.

Hence the TEV protease concentration of 1µg/µl was used to cleave the whole eluted protein fraction. The TEV protease was added in required amount to the whole eluted protein fraction, and allowed to incubate for 5 hours at room temperature.

The cleaved protein fraction was then incubated with 10 ml of the nickel beads for 30 minutes, 30 mM of imidazole was added to it and the whole mixture was poured into a column, where the required protein was collected as a flow through and the nickel beads along with the attached histidine tag was retained in the column.

The gel electrophoresis of all the fractions collected at every step was carried out as shown in Figure 14.5.

Figure 14.5
Gel electrophoresis of the eluted protein fractions in different stages

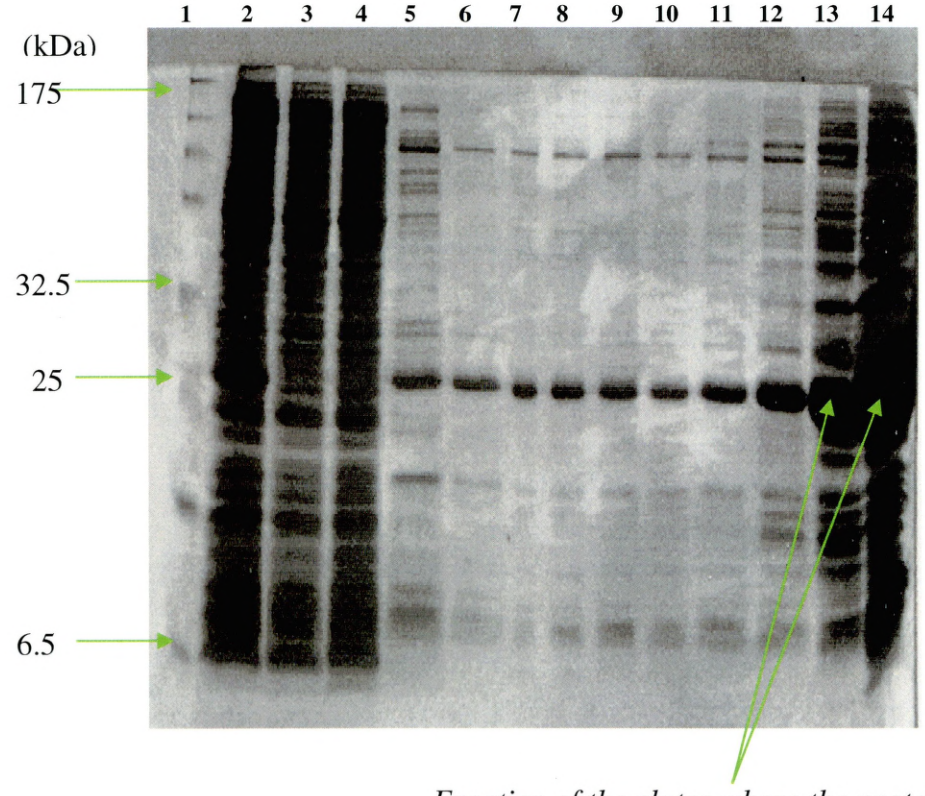


Figure 14.5 is the gel electrophoresis of the eluted protein fraction at different stages. Lane 1 is the protein marker, lane 2 is the supernatant before binding with the nickel beads, lane 3 is the supernatant after binding with the nickel beads (proteins that initially failed to bind to the nickel column), lane 4 is the first wash, lane 5 is the second wash, lane 6 is the third wash with the wash buffer of the nickel column to remove any unbound protein. The lanes 7 to 14 are the different stages of the eluted UBE2J1 protein from the nickel column by passing through the elution buffer through the nickel column.

Figure 14.6 Gel electrophoresis of the TEV protease cleaved protein sample

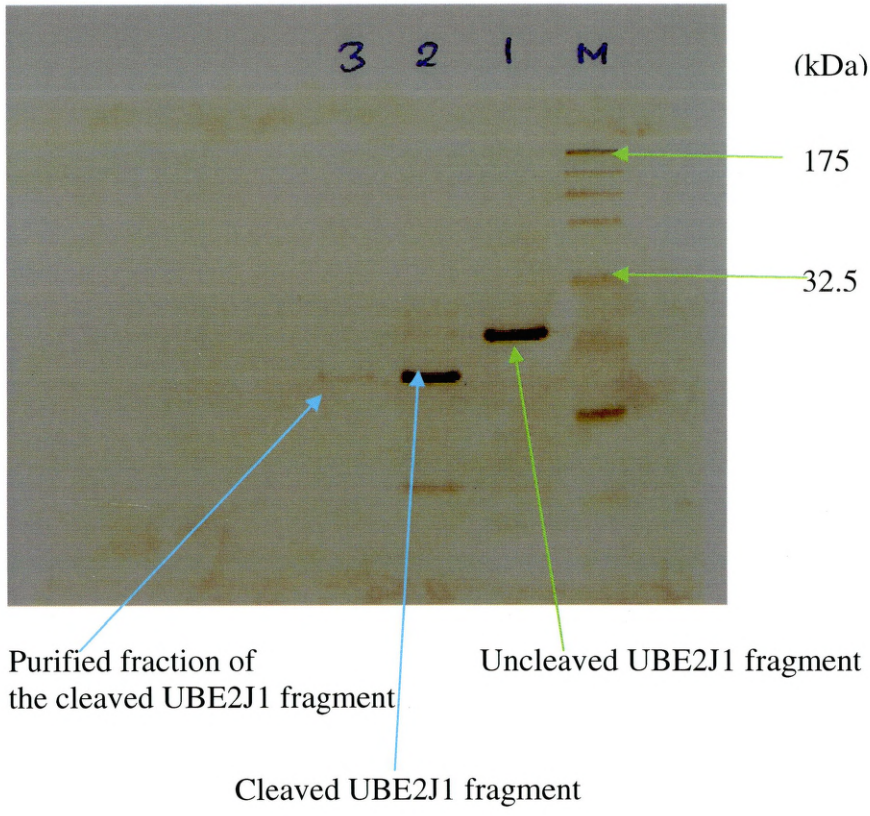


Figure 14.6 is the gel electrophoresis of the uncleaved fraction in lane 1, TEV protease cleaved UBE2J1 fraction in lane 2, and the purified UBE2J1 protein fraction in lane 3.

14.1.4. Gel purification (by gel permeation chromatography)

The protein was further purified by gel permeation chromatography. The buffer used as detailed in the methods section 9.24. The chromatograph obtained from the gel permeation chromatography, is as shown in Figure 14.7.

Figure 14.7 The chromatogram of gel permeation chromatography

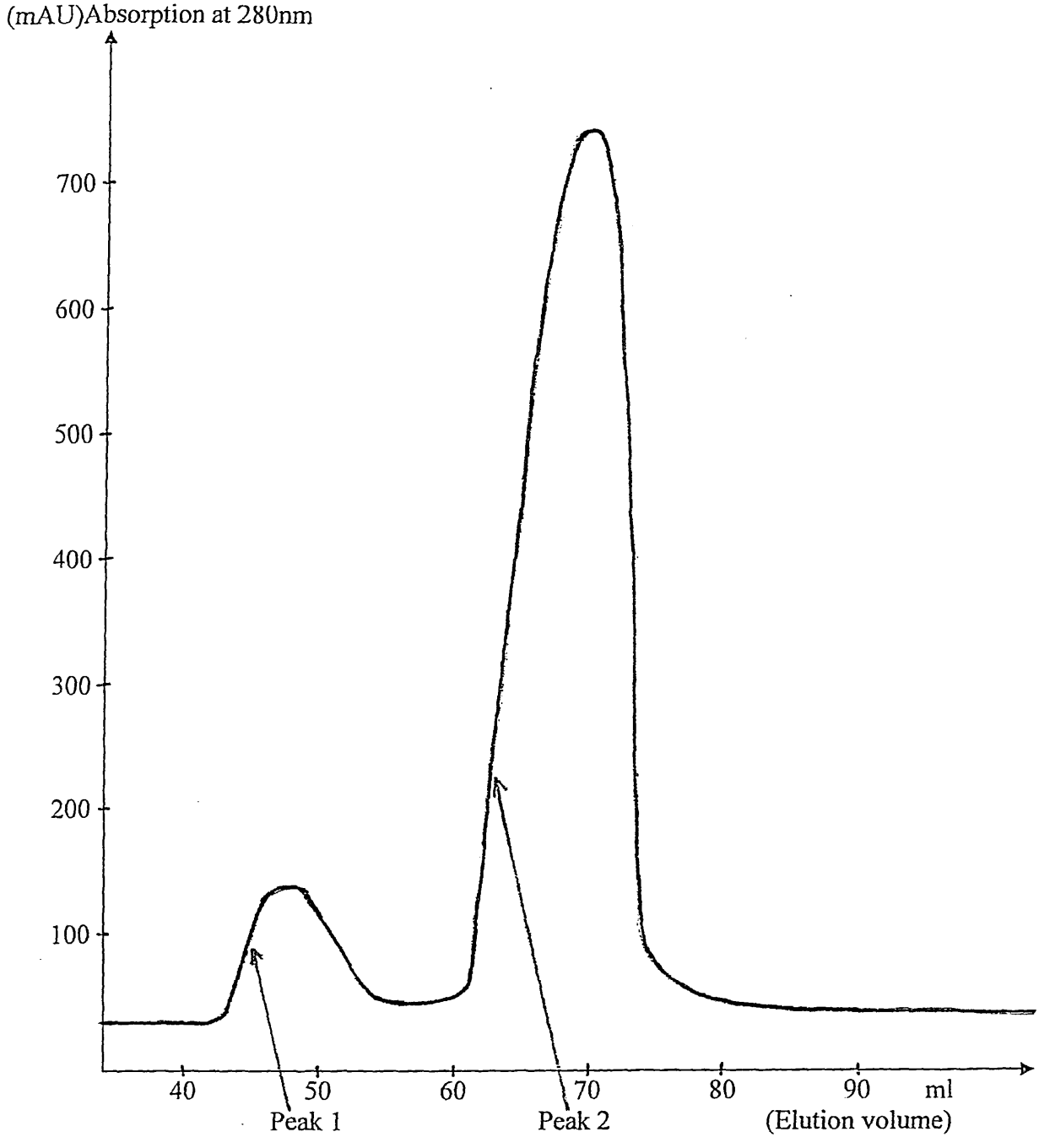
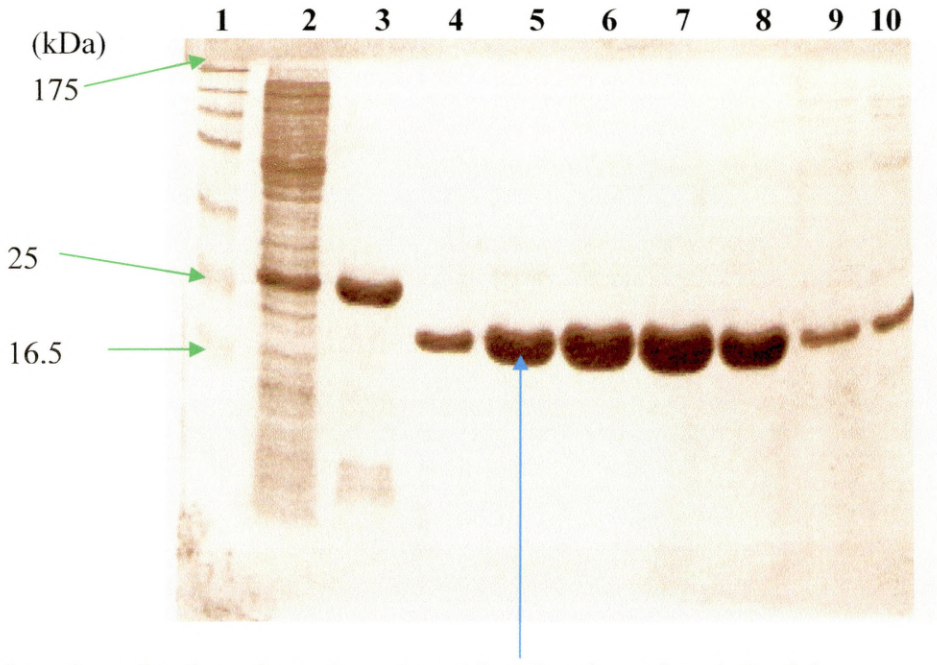


Figure 14.7 is the result of the gel purification of the protein sample collected in 2ml fractions. There are 2 peaks where peak 1 is the dimerised form of the protein UBE2J1 fragment. Peak 2 is the monomeric form of UBE2J1 protein; the fractions of the purified protein under peak 2 were collected for crystallization.

As can be seen, there are 2 peaks, among which the main protein are in the fractions under the second big peak. All the fractions of peak 2 were collected separately and gel electrophoresis was carried out before pooling them.
The gel electrophoresis result is as illustrated in Figure 14.8.

Figure 14.8 Gel electrophoresis of gel purified fractions



Starting from lane 5 to lane 8 are the gel purified fraction of peak 2 of the chromatogram shown in figure 14.9.

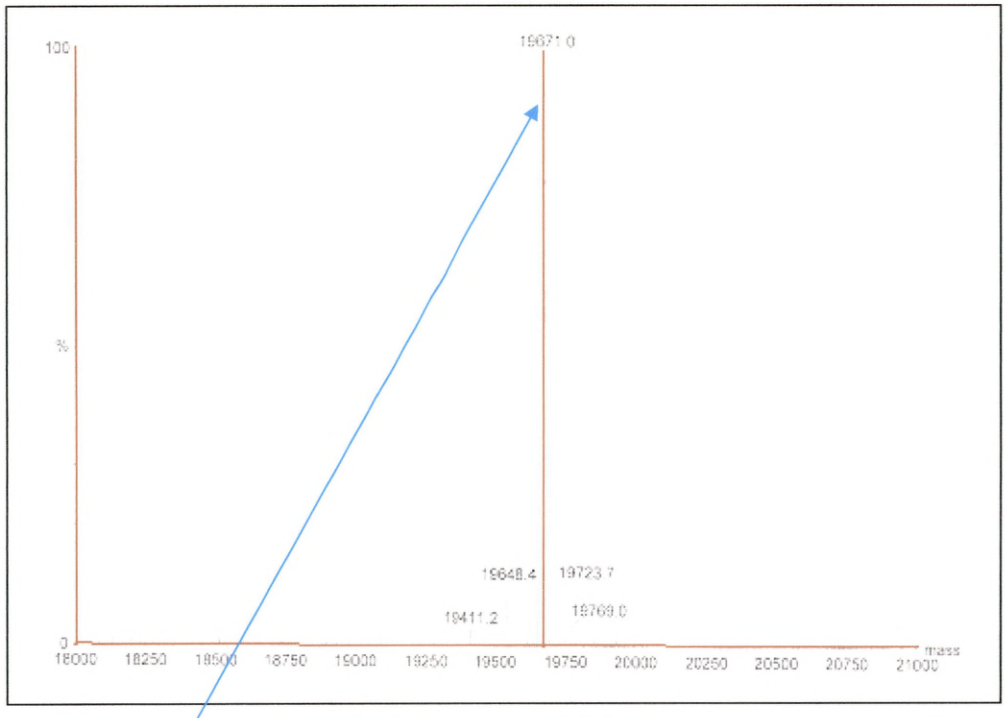
Figure 14.8 is the gel electrophoresis of the gel purified fractions. Lane 1 is the protein marker, lane 2 is the unbound nickel column supernatant. Lane 3 is the eluted protein without the TEV protease digestion, lane 4 is the TEV protease cleaved protein which is eluted from the nickel column. Lanes 5 to 8 are the different gel purified fraction of the big peak in the gel purification chromatogram as shown Figure 14.7 and lanes 9 & 10 are the fractions of the first small peak, which is the dimerised form of UBE2J1.

As can be seen from the gel bands, the majority of the protein is retained in the second big peak and as there are no impurity bands for these four fractions, it was considered best to pool the four fraction of the second big peak.

14.1.5. Mass spectrometry

To further confirm that only the required protein was present in the pooled fraction from the gel purified sample and no other contaminated proteins were present, the mass spectrometric analysis was carried out. The sample was given in a volume of 20 µl whose concentration was 20pmol/µl. The result of the mass spectrometric analysis is as shown in Figure 14.9.

Figure 14.9 Mass spectrometric result



UBE2J1 fragment

Figure 14.9 shows the mass spectrometric analysis result, from which it can be said that there is only one protein of the molecular weight of 19671.0 Daltons, which is the correct size for the UBE2J1 fragment.

Figure 14.10 A suspected crystal growth in one of the crystal trial plates

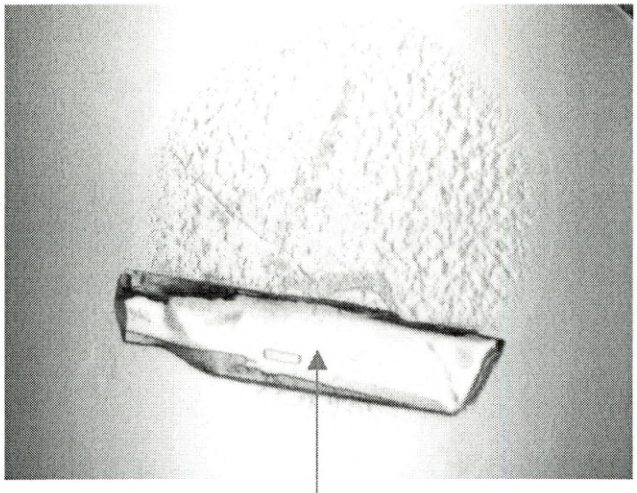


Figure 14.10 is a crystal formation in one of the crystal trial plates set for crystallization, picture taken by the crystallography machine.

Below is the diffraction pattern obtained of the salt crystal and the actual diffraction pattern of a protein as an example, so as to show how a diffraction pattern of a salt crystal is different from that of the diffraction pattern of a protein.

Figure 14.11.1
Diffraction pattern of the crystal obtained

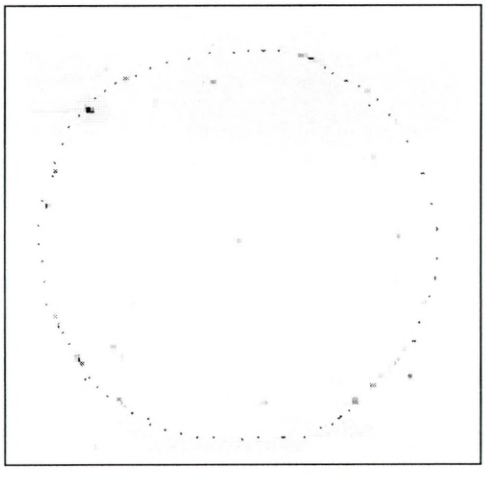


Figure 14.11.2
An example of a typical diffraction pattern of a protein crystal

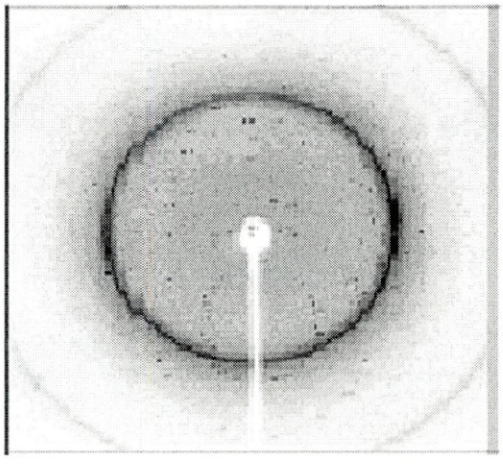


Figure 14.11.1 is the diffraction pattern of the crystal that was obtained which actually is a salt crystal (ammonium sulphate crystal); and to the right of it is Figure 14.11.2 which is an example of a typical diffraction pattern of a protein crystal, so as to give a visual comparison of the two diffraction patterns.

14.2. Results of Circular Dichroism (CD) spectrum analysis

After four months of incubation under all conditions (see appendix), no crystal was obtained for the UBE2J1 protein fragment. In order to try and understand why this UBE2J1 fragment did not crystallise, a Circular Dichroism (CD) spectroscopic analysis was carried out, to get more information of the secondary structural composition of the UBE2J1 fragment that had been unsuccessful to crystallise.

The CD spectrum analysis was carried out for both the far UV and the near UV.

The CD spectrum obtained for the far UV ranging from 190 to 260 nm is as below:

Figure 14.12 Far UV CD spectrum of the protein UBE2J1

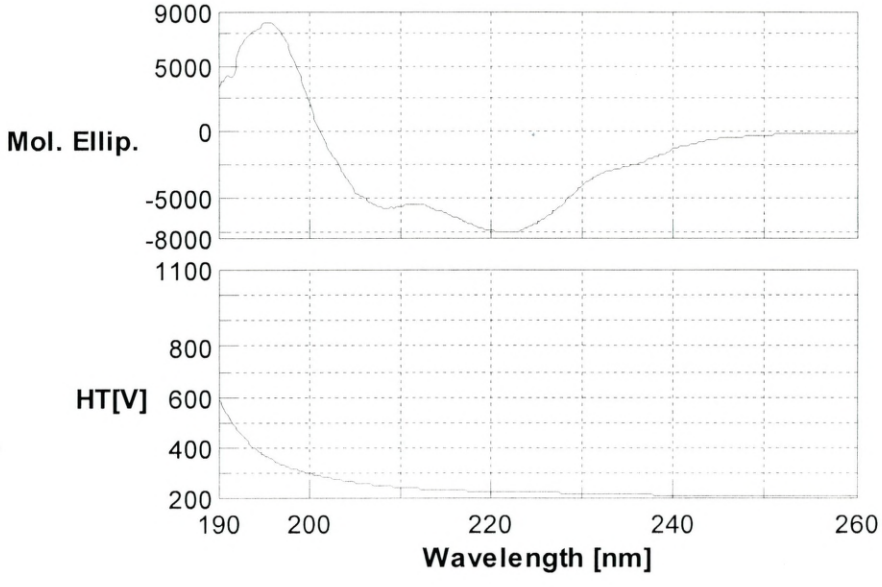


Figure 14.12 is the graphical representation illustrating the spectrum of the UBE2J1 fragment in the far UV region. (Mol. Ellip.) as seen in the Y-axis, represents molar ellipticity, and HT[V] represents high tension voltage. The upper panel shows the CD spectrum and the lower panel shows the corresponding high tension voltage.

Different CD instruments have different HT operating values, like the J-600 has a maximum HT value of 574 (with a fluctuation of 5) V. The J715 has a maximum value of 587V (with a fluctuation of 5). Hence for these instruments around 600V is considered maximum safe operating HT voltage.

Molar ellipticity (θ)m is measured in degrees decilitres mol^{-1} $decimeter^{-1}$.

The CD spectrum obtained for the near UV range from 250 to 320nm is as follows:

Figure 14.13 Near UV CD of the protein UBE2J1

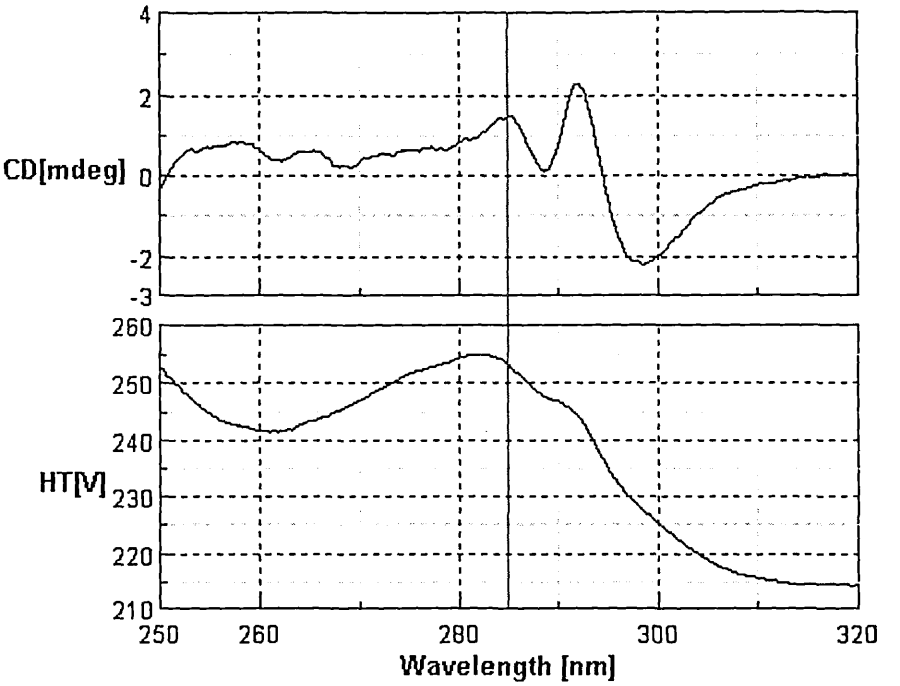


Figure 14.13 is the graphical representation illustrating the spectrum of the UBE2J1 fragment in the near UV region.

These two spectral results obtained above were used in the DICHROWEB server for further structural analysis.

The CD spectral data was then analysed with the help of "DICHROWEB" an online analysis server where the secondary structural properties were estimated. The result obtained was the secondary structural content of the protein fragment UBE2J2. The output of the analysis obtained is as stated below:

Table 14.2 Result obtained from Dichroweb of the secondary structural component, which has been calculated from the CD spectrum of Figure 14.15

NRMSD:0.031

Helix segments per 100 residues: 2.793

Strand segments per 100 residues: 3.948

Ave helix length per segment: 7.640

Ave strand length per segment: 5.284

Helix1	Helix2	Strand1	Strand2	Turns	Unordered	Total
0.10	0.11	0.13	0.08	0.18	0.39	0.99

From the results of the DICHROWEB, that had analysed the CD spectrum which had been obtained, the sum total of the two fractions of each of helix and sheet were, the protein contained around 21% helix (sum of Helix1 and Helix2), 21% sheet (sum of strand1 and strand2), 18% turns and 39-40% unordered.

The normalised root mean square deviation (NRMSD) obtained here was 0.031. The different percentages of the secondary structure content were calculated from the CD spectral data by the Dichroweb as shown in Table 14.2. The NRMSD parameter is actually a measure of how well the theoretical CD spectrum calculated from the derived secondary structure composition matches the experimental data over the entire length of wavelength of interest. A low value of the NRMSD (<0.1) considers a good result, but it does not mean that it is accurate. As the NRMSD result obtained in this experiment was 0.031, it was considered to be a good result. The fact that there was a particular percentage of secondary structure, adds to the correctness of the UBE2J1 that was being used to crystallise. Hence from the far UV CD spectrum the secondary structural properties of the protein UBE2J1 were derived.

The near UV spectral analysis result was used to monitor the tertiary structure of the protein. The spectral bands in this region of 240 – 320 nm wavelength were of the aromatic residues, like mostly Tryptophan. The spectral bands in the wavelength of 240 to 320nm were of the aromatic residues, which are very sensitive to the tertiary structural changes. Tryptophan is the only aromatic residue that has a significant contribution to the near UV CD spectrum analysis. There are two tryptophan residues around the active site residue Cysteine of the UBE2J1 fragment. A list of percentage of the amino acids in the protein UBE2J1 as given by the CD spectral analysis by the Dichroweb server is given in the appendix section.

From the circular dichroism spectral analysis it was concluded that the fragment of the protein UBE2J1 that was tried to crystallise, had the secondary structural contents of around 21% each of α-helices and β-sheets, but around 40% of the peptide sequence was unordered. This could be the possible reason why the fragment of UBE2J1 did not crystallise. The disadvantage in CD spectral analysis is that it does not provide information that which residues are disordered. Hence other methods have been tried that would predict the globular part and the disordered part in the peptide sequence.

14.3. Conclusion and future research

From the UV CD spectral analysis it was concluded that, although the protein contained around 21% helix, 21% sheet and 18% turns, there is a 39-40% unordered region in the protein. Hence this unordered region contributes to the improper folding of the protein. This required to design a new set of primers by taking a different domain or fragment of the protein, but due to time and resource constraint, it was not possible to carry out the analysis again. Some ideas have been put forward for future research to crystallise the same UBE2J1 protein, considering the globular and disordered regions in the protein. The programs that predict the globularity and disorder in the peptide sequence are discussed as follows:

14.4. To check the globularity and disorder domains in the UBE2J1 fragment

To check whether the protein UBE2J1 has any predicted "globular" and "disordered" domains, tools like GLOBPLOT and RONN were used.

Another technique that is often used, is limited trypsin proteolysis; after running the trypsin treated protein on SDS-PAGE, any bands containing the core domains (which tend to be more stable to proteolysis) can be analysed by mass spectrometry and N-terminal sequencing to ascertain which region of the protein they correspond to. This information could be then used to clone that region.

The other problem may be with protein aggregation - this can be checked by light scattering or by running a native gel. It would give an indication of how homogeneous the protein sample is, which, is an important factor for the crystallization of a protein.

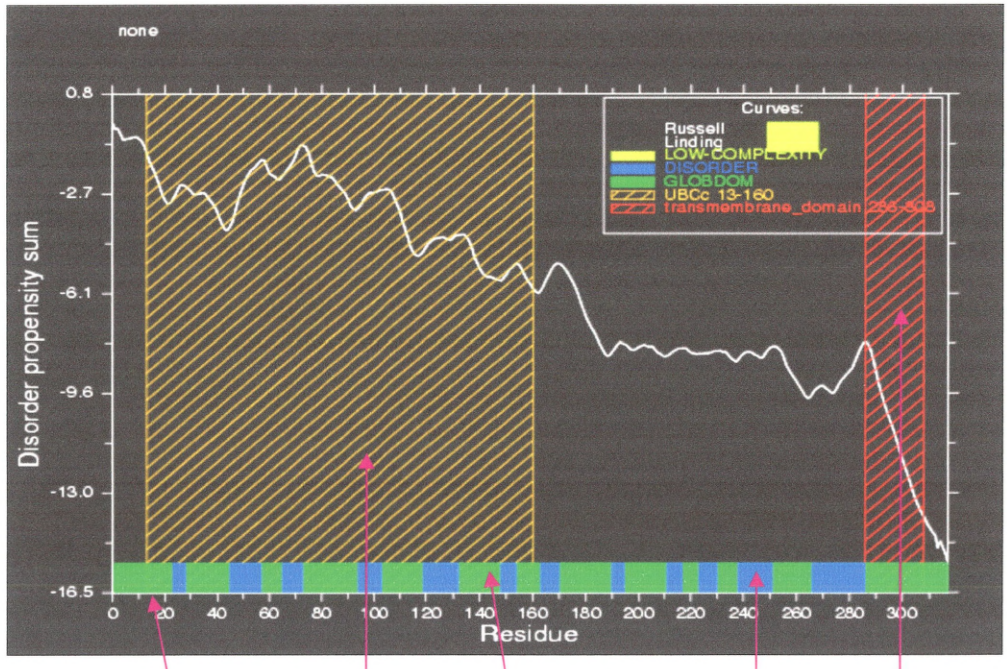
Purity is the other important reason. It has been found that even if a protein looks clean on SDS PAGE, a gel filtration polishing step at the end can make the difference between a protein crystallizing and not crystallizing.

14.4.1. Globplot

GLOBPLOT is a web based tool accessible at (http://globplot.embl.de), used to find out the disorder and globular segments of a protein. It is known that many functionally important segments of a protein lie outside the globular domain that is classified as disordered. Globplot is a computational tool which helps to find out domains of order or

disorder in a peptide sequence. The disorder of a protein may be termed as a lack of regular secondary structures or it could be that the protein is not properly folded. The result of the globplot analysis is as shown in Figure 14.16. The resulting Globularity/ non-globularity is named Russell/Linding, which combined both the random coil and secondary structure of the protein (Linding et al., 2003; Linding, 2004).

Figure 14.14 Result of globplot



X-axis represents the whole length of amino acids

This region of amino acids has been recognised as UBCs

Globular region

Disordered region

Transmembrane domain region

Figure 14.14 is a graphical representation of the ordered or the disordered region of the whole protein UBE2J1. The region marked in green is the globular (ordered) region, whereas the region marked in blue is the disordered region. The transmembrane domain region is shaded in red and the region shaded in yellow is recognised by the Globplot server as the UBCs. The globular and disordered region is further illustrated along the peptide sequence as in Table 14.3.

Table 14.3 is a interpretation of the graph in Figure 14.17, predicting the ordered and disordered regions of UBE2J1 by Globplot, by Russell/Linding.

Table 14.3
Globular and disordered region of UBE2J1 protein

>none_Disorder 23-28, 45-57, 65-73, 94-103, 119-132, 148-154, 163-170, 190-195,211-217,223-230,238-251,266-286
metrynlksp avkrlmkeaa el**KDPTDH**yh aqplednlfe whft**VRGPPD SDFDGGV**yhg rivl**PPEYPM KPP**siillta ngrfevgkki clsl**ISGHHPE TWQ**pswsirt allaiigf**MP TKGEGAIGSL DY**tpeerral akksqdf**CCE GCGS**amkdvl lp**LKSGSDSS** qadqeakela rqisfkaev**N SSGKT**isesd lnhsfsltdl **QDDIPTT**fqg at**ASTSYGLQ** nssaasf**HQP TQPVAKNTSM S**prqrraqqq sqrrl**STSPD VIQGHQPRDN HTDHGG**savl iviltlalaa lifrriylan eyifdfel

The result shown in Table 14.3, shows the ordered and the disordered regions of the whole protein UBE2J1. The amino acids in lower case are the globular amino acids, whereas those that are in uppercase and in bold are the disordered amino acids.

From the result as shown above it is clear that the active site residue Cysteine (shown as "C" above) and some residues around it is globular, but there are also some amino acids that are disordered.

This is further supported by the GLOBPLOT of the recently crystallised protein UBE2J2 as shown below:

Table 14.4 Globular and disordered regions of the protein UBE2J2

>none_Disorder 1-9, 35-39, 50-61, 70-76, 82-87, 99-107, 125-131, 178-185
MSSTSSKRAp ttatqrlkqd ylrikkdpvp yica**EPLPS**n ilewhyvvr**G PEMTPYEGGY Y**hgklifpr**E FPFKPP**siym i**TPNGRF**kcn trlclsit**DF HPDTWNP**aws vstiltglls fmve**KGPTLG S**ietsdftkr qlavqslafn lkdkvfcelf pevveeikqk qkaqdel**SSR PQTLP**

Shown in Table 14.4 are the ordered and the disordered region of the fragment of UBE2J2 that had recently been crystallised. The amino acids in lower case are the globular proteins, whereas those that are in uppercase and in bold are the disordered amino acids.

As can been seen in Table 14.4, UBE2J2 also have the active site residue cysteine (shown as "C" above) and some of the other amino acids in the globular region, but there are some amino acids that are disordered as well.

Though GLOBPLOT is one of the most recommended tools to predict regions of globularity and disorder in the protein, here it can be seen from the GLOBPLOT analysis of both the proteins UBE2J1 and UBE2J2, that there is not much significant difference between the two UBE2J1 and UBE2J2 disorder prediction. Both show almost the same regions of globularity and disorder; although very recently UBE2J2 has been crystallised.

14.4.2. RONN: to predict the disordered region of a protein

RONN is another tool for predicting the disorder in a protein (http://www.strubi.ox.ac.uk/cgi-bin/disorder_results_jan2005.cgi). The result is a graphical representation showing the ordered and disordered regions of the protein (Yang et al., 2005)

Figure 14.15 Disorder prediction of UBE2J1 by RONN prediction server

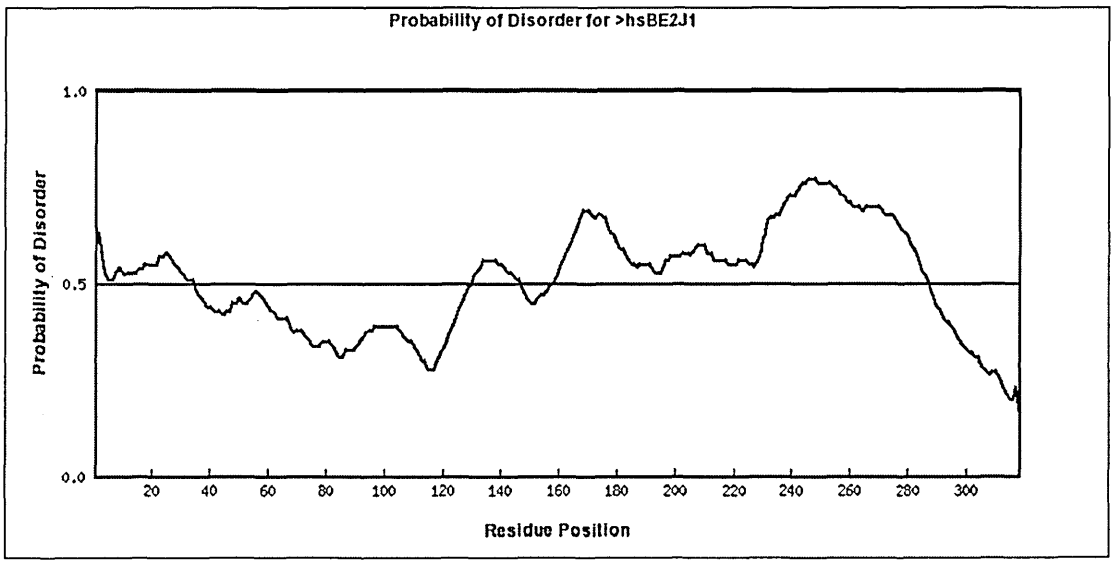


Figure 14.15 is the graph of the globularity and disordered region of the protein UBE2J1. The region below the line marked 0.5 is the globular region, whereas the region above the threshold value is the disordered region of the protein UBE2J1.

This graph has been interpreted on the amino acid level as shown in table 14.5, to show the globular and disordered region of the protein UBE2J1.

Table 14.5
Continuation of the disorder prediction of UBE2J1 by RONN prediction server

```
1           METRYNLKSP AVKRLMKEAA ELKDPTDHYH AQPLEDNLFE WHFTVRGPPD  50
            ********** ********** ********** ****_____ _________
51          SDFDGGVYHG RIVLPPEYPM KPPSIILLTA NGRFEVGKKI CLSISGHHPE 100
            _________ __________ __________ __________ __________
101         TWQPSWSIRT ALLAIIGFMP TKGEGAIGSL DYTPEERRAL AKKSQDFCCE 150
                                             ********** ******____
151         GCGSAMKDVL LPLKSGSDSS QADQEAKELA RQISFKAEVN SSGKTISESD 200
            ________ ** ********** ********** ********** **********
201         LNHSFSLTDL QDDIPTTFQG ATASTSYGLQ NSSAASFHQP TQPVAKNTSM 250
            ********** ********** ********** ********** **********
251         SPRQRRAQQQ SQRRLSTSPD VIQGHQPRDN HTDHGGSAVL IVILTLALAA 300
            ********** ********** ********** ******____ _________
301         LIFRRIYLAN EYIFDFEL
            _________ ________
```

From the result of the disorder prediction by RONN, it is found that the disordered regions are 1- 34, 131- 146, 159- 286. In this RONN analysis tool, the cut off limit is 0.5, where any amino acid having a probability of disorder value greater than 0.5 is considered to be disorder. Hence it can be seen that most of the region of the protein UBE2J1 is disordered except for around 100 amino acid residues in the region of amino acids ranging from 35 – 130, which also consists the active site residue Cysteine.

Figure 14.16 Disorder prediction of UBE2J2 by RONN prediction server

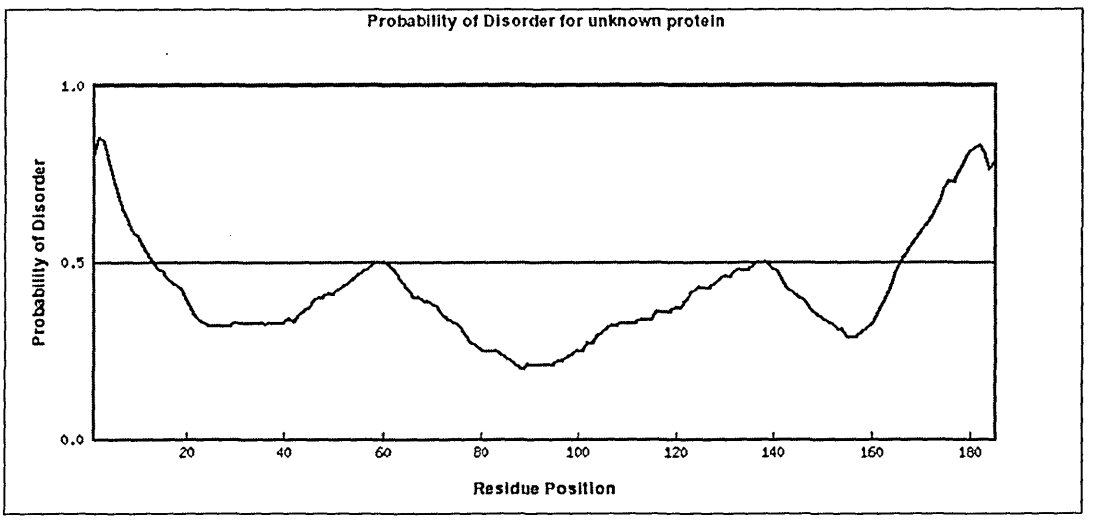


Figure 14.16 is the graph of the globularity and disordered region of the protein UBE2J2. The region below the line marked 0.5 is the globular region, whereas the region above the threshold value is the disordered region of the protein UBE2J2.

The region of disorder and of the globular part of the protein UBE2J2 whose graphical representation is as shown in the previous page in figure 14.16, is shown below at the amino acid level.

Table 14.6
Continuation of the disorder prediction of UBE2J1 by RONN predicion server

```
1           MSSTSSKRAP TTATQRLKQD YLRIKKDPVP YICAEPLPSN ILEWHYVVRG  50
            ********** **_______ _______ _______ _______
51          PEMTPYEGGY YHGKLIFPRE FPFKPPSIYM ITPNGRFKCN TRLCLSITDF  100
            _______ _______ _______ _______ _______ _______
101         HPDTWNPAWS VSTILTGLLS FMVEKGPTLG SIETSDFTKR QLAVQSLAFN  150
            _______ _______ _______ _______ _______
151         LKDKVFCELF PEVVEEIKQK QKAQDELSSR PQTLP
                    _____**** ********** *****
```

From the RONN result of the protein UBE2J2 above, it is very clear that UBE2J2 as a whole is a globular protein, with the majority of the portion of the protein is globular. A very minor portion of disordered is between amino acid 1 - 12, and 167 – 185.

From the RONN result of UBE2J1, various regions of the protein could be used to design primers, and try to crystallise them. Short fragments of the UBE2J1, like amino acid 35- 130, or 35 – 158, could be the used to design primers.

14.4.3. Trypsin proteolysis

The other method is to do a trypsin digest of the protein and try to find out the regions of the protein that are not folded or loosely hanging on the surface. By the trypsin proteolysis, the trypsin cleaves at the amino acids lysine-X and agrinine-X bond. These amino acids are abundantly available in the protein, but the trypsin proteolysis is only possible on the surface region of the protein, as it cannot penetrate at the compact core region. Trypsin cleaves at the particular sites, and this trypsin digested protein is then run on the SDS PAGE to separate all the cleaved regions. Each region can be cut out from the gel bands, protein extracted and either sequenced or done a mass spectrometric analysis, to find out the regions that are in the core domain. That core domain fragment could then be considered for crystallisation (Beynon and Bond, 1989).

14.4.4. Protein purification

The final stage of protein preparation is to purify the protein UBE2J1. An affinity chromatography could be carried out from ubiquitin-affinity-gel followed by the size separation (Gel filtration) of the protein, or an alternative purification procedure could be carried out to separate the protein by cation exchange chromatography followed by gel permeation chromatography.

CHAPTER FIFTEEN

RESULTS & DISCUSSION OF COMPUTATIONAL STRUCTURE PREDICTION (HOMOLOGY MODELLING)

15.1 Results and discussion of computational structure prediction

15.1.1. Modelling, using DeepView

The modelling steps in DeepView were as follows:

Model Building: To generate the backbone of the model, the templates are evaluated by their sequence similarity to the target sequence, where significantly deviating atom positions are excluded. The template coordinates are then used to model the regions of insertions and deletions in the target-template alignment. These parts are generated by the "constraint space programming" (CSP). The best loop is selected using the scoring scheme, which looks for favourable interactions like hydrogen bond formation, steric hindrance, and force field energy.

Side chain modelling: The side chain modelling is also based on the evaluation of the corresponding residues in the template structures. The model side chains of the conserved residues, are built by replacing the template structure side chains. The backbone dependent rotamer library gives a selection of the possible side chain conformations.

Energy minimizations: The protein model is regularised in the last step, by the steepest descent energy minimization, using the GROMOS96 force field (Schwede et al., 2003).

15.1.2. Obtaining the model of UBE2J1 by DeepView

The domain information of the protein sequence (UBE2J1) was needed to be found first before homology modelling was carried out. There are various tools for protein domain search like the PfamA, InterPro, ProDom, SBASE. The tool that was used here in the analysis was PfamA, and the result of the domain search is in Figure 15.1.

Figure 15.1 Results obtained from Pfam domain prediction

Model	Seq-from	Seq-to	Score	E-value	Alignment	Description
UQ_con	14	149	95.8	1.2e-25	global	Ubiquitin-conjugating enzyme

Figure 15.1 shows the range of the UBE2J1 peptide sequence that was found suitable for carrying out homology modelling. What Pfam had done is, it had carried out a multiple sequence alignment of the UBE2J1 peptide sequence with all UBCs and had predicted the most conserved regions in the alignment, which is from amino acid 14 to 149.

It was found that the sequence region of amino acid 14 to amino acid 149 was the best region, as it had the greatest homology, with a significant E-value, which is < 0.05. Hence the peptide sequence around that region was taken to do the homology modelling.

E-value is the Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. Lower the E value, the more significant the score. The S score is a measure of the similarity of the query to the sequence shown. The E-value is a measure of the reliability of the S score

(http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html).

The actual equation is $E=Kmn(e^{-\lambda S})$, where K and λ represent natural scales for the search space and the scoring system respectively. The rest of the equation represents the size of the query (m), the size of the database (n), and of course the S score.

The typical threshold for a good E-value from a BLAST search is $e^{-5}=(10^{-5})$ or lower.

The reason for such low values is that an E=0.001 in a million entry database would still leave 1000 entries due to chance. An $E=e^{-6}$ would only leave one entry due to chance (http://www.osc.edu/research/bioinformatics/FAQ/evalue.shtml).

The secondary structural components of the protein UBE2J1 was required to be predicted, which was carried out by the GOR4 secondary structure prediction server (Garnier, Gibrat and Robson, 1996) as shown in Figure 15.2.

Figure 15.2 Secondary structural contents of UBE2J1

```
         10         20         30         40         50         60         70
          |          |          |          |          |          |          |
METRYNLKSPAVKRLMKEAAELKDPTDHYHAQPLEDNLFEWHFTVRGPPDSDFDGGVYHGRIVLPPEYPM
ccceecccchhhhhhhhhhhhhcccccccccccccccceeeeeeeecccccccccccceeeeeeecccccc
KPPSIILLTANGRFEVGKKICLSISGHHPETWQPSWSIRTALLAIIGFMPTKGEGAIGSLDYTPEERRAL
ccceeeeeeccccecccceeeeeeecccccccccchhhhhhhhhhhhcccccccccceeeecccchhhhhhh
AKKSQDFCCEGCGSAMKDVLLPLKSGSDSSQADQEAKELARQISFKAEVNSSGKTISESDLNHSFSLTDL
hhhcccceeeecccccccceeeeecccccccchhhhhhhhhhhhhhhhhhhcccccccecccccccccceecc
QDDIPTTFQGATASTSYGLQNSSAASFHQPTQPVAKNTSMSPRQRRAQQQSQRRLSTSPDVIQGHQPRDN
ccccccceecccccceecccccccccccccccceecccccchhhhhhhhhhhcccccccceeecccccc
HTDHGGSAVLIVILTLALAALIFRRIYLANEYIFDFEL
cccccccceeeehhhhhhhhhhhhhhhhhhhhcceeceec

Sequence length :   318


   Alpha helix      (Hh) :    84 is  26.42%
   3₁₀ helix        (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :    66 is  20.75%
   Beta turn        (Tt) :     0 is   0.00%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :   168 is  52.83%
   Ambigous states  (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```

Figure 15.2 shows the secondary structural component of the peptide sequence UBE2J1 obtained from GOR4 (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html).

The knowledge of the secondary structural components was necessary when selecting a range of amino acids for modelling. This gave an idea of the secondary structural components, so that when selecting a range of amino acids for modelling, a complete range of the secondary structural component is selected.

A template that was found to be the closest homologue of UBE2J1 peptide sequence, was the recently found UBE2J2 (pdb file-2F4WB) crystal structure.

This template (2F4WB) was then retrieved in Deepview from the PDB.

In the structure of 2F4WB, there is a gap after leucine (which is just after the active site residue cysteine), where 9 amino acids are missing. It is often the case in structures solved by X-ray crystallography that there are residues that cannot be modelled due to their high mobility or disorder. In 2F4WB the amino acids are indeed present in the crystal, but no electron density has been observed. The pairwise alignment of UBE2J1 with the template 2F4WB is as in Figure 15.3.

Figure 15.3 Alignment of UBE2J1 with the template (2F4WB)

```
# Identity:    60/139 (43.2%)
# Similarity:  90/139 (64.7%)
# Gaps:        6/139 ( 4.3%)

hsBE2J1: rlmkeaael-kdptdhyhaqplednlfewhftvrgppdsdfdggvyhgri
         ||.::...: |||..:..|:||..|:.|||:.||||..:.::||.|||::
2F4WB  : rlkqdylrikkdpvpyicaeplpsnilewhyvvrgpemtpyeggyyhgkl

hsBE2J1: vlppeypmkppsiilltangrfevgkkiclsisghhpetwqpswsirtal
         :.|.|:|.|||||.::|.||||:...::||||:..||:||.|:||:.|.|
2F4WB  : ifprefpfkppsiymitpngrfkcntrlclsitdfhpdtwnpawsvstil

hsBE2J1: laiigfmptkge--gaigsldytpeerralakksqdfcc
         ..::.||..||.  |:|.::.|:|   :|.||.:|..|..
2F4WB  : tgllsfmvekgptlgsietsdft---krqlavqslafnl
```

Figure 15.3 shows the pairwise alignment of UBE2J1 with the template 2F4WB. The percentage identity is around 43%.

The aligned UBE2J1 and the template have the percentage identity of around 43% and similarity of around 65%. Theoretically 30% sequence identity is the lowest limit to carry out homology modelling. The pairwise alignment of these two sequences were carried out in "L-align" (http://www.ebi.ac.uk/emboss/align/index.html), to try to increase the percentage identity by aligning various segments of the sequences, but it was found that although some trials gave an increased percentage identity, there were large gaps introduced. Hence the alignment that was generated by the DeepView software was taken to carry out the homology modelling. Shown in Figure 15.4 is the pairwise sequence alignment of UBE2J1 with that of the template (2F4WB), generated by DeepView.

Figure 15.4 Pairwise sequence alignment generated by DeepView

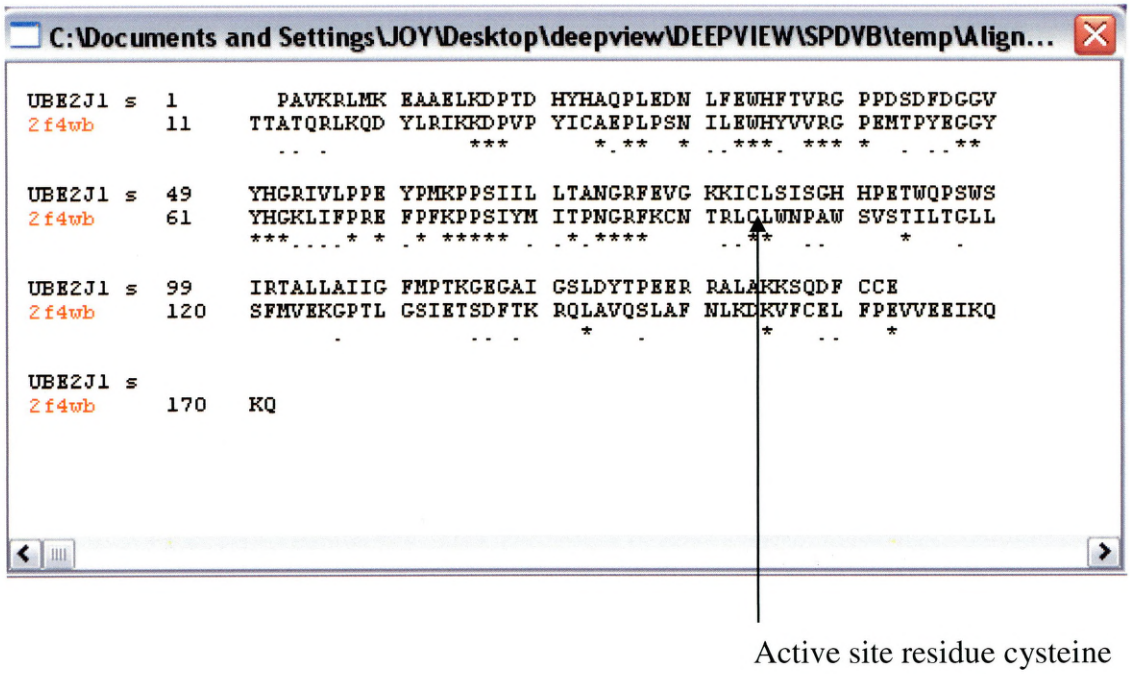


Figure 15.4 is an alignment generated by Deepview. 2F4WB is the template based on which the query sequence UBE2J1 was modelled.

In the template UBE2J2 (2F4WB), there are 9 amino acids missing. Although the actual protein has those 9 amino acids, no diffraction patterns have been obtained from the UBE2J2 crystal. Hence gaps were then introduced in the peptide sequence of the template, in the region were there are 9 amino acids missing when carrying out the homology modelling, so that the model of UBE2J1 will also have that region highlighted. Finally the alignment had been submitted to DeepView to generate the model is as in Figure 15.5.

Figure 15.5 The alignment that was generated by DeepView and the gap introduced manually at the region of missing 9 amino acids

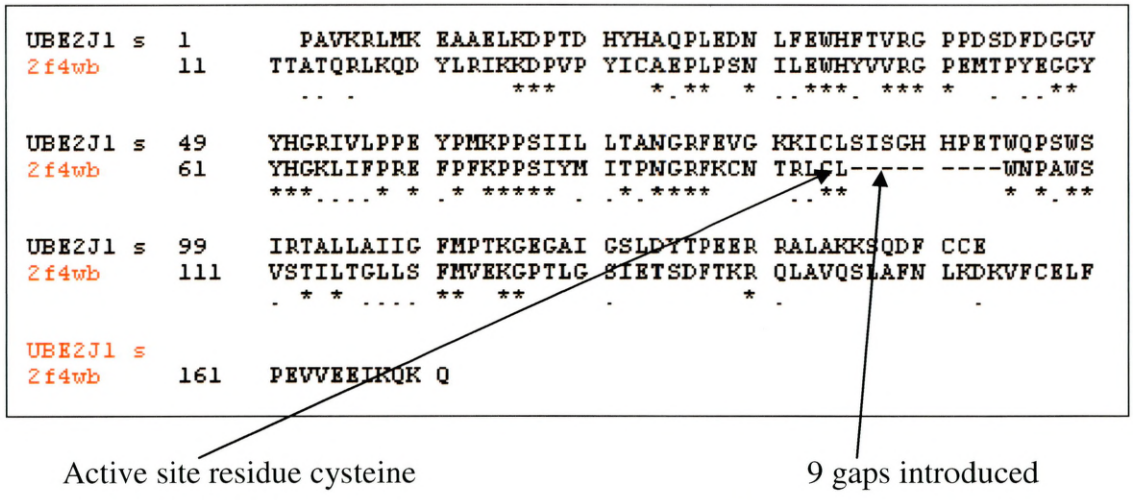


Figure 15.5 is an alignment generated by Deepview. 2F4WB is the template based on which the query sequence UBE2J1 was modelled. 9 gaps were introduced manually in the 2F4WB sequence in the DeepView alignment as there are 9 amino acids missing in the structure of 2F4WB in the region arrow marked.

This aligned sequence of UBE2J1 was then sent to SwissPDB to be modelled. Stated as follows, is the discussion of the model of UBE2J1 that was obtained from Swiss Model.

15.1.3. Evaluating and optimising the model sent by SWISS PDB

The model of UBE2J1 received from the Swiss Model, along with a geometric check report (What Check report) was viewed in DeepView. The modelled UBE2J1 obtained from the SwissPDB was first coloured by B-factor to look for high or low B-factor values or to identify areas of structural uncertainty. The model of UBE2J1 colouring by B-factor is as in Figure 15.6.

15.1.3.1. Colour by B-factor

Figure 15.6 Model of UBE2J1 coloured by B-factor

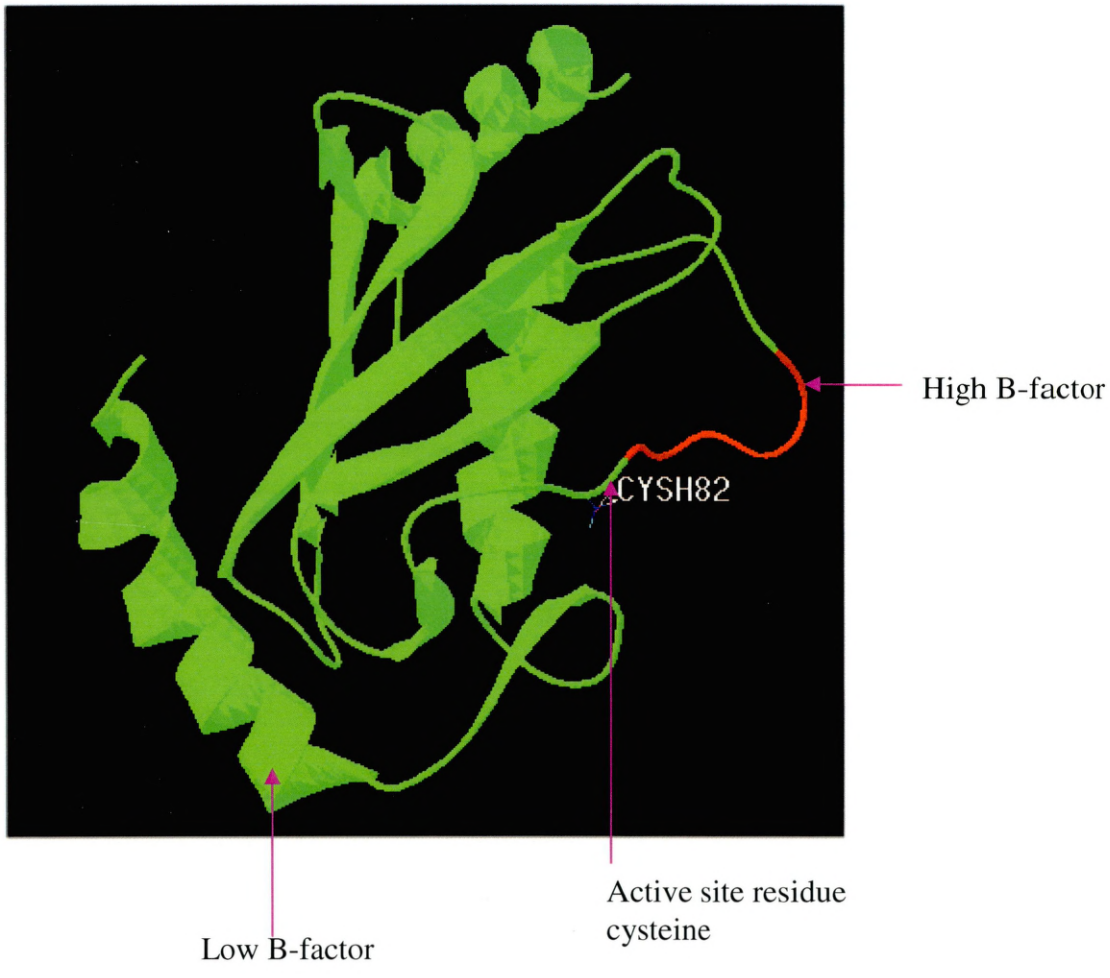


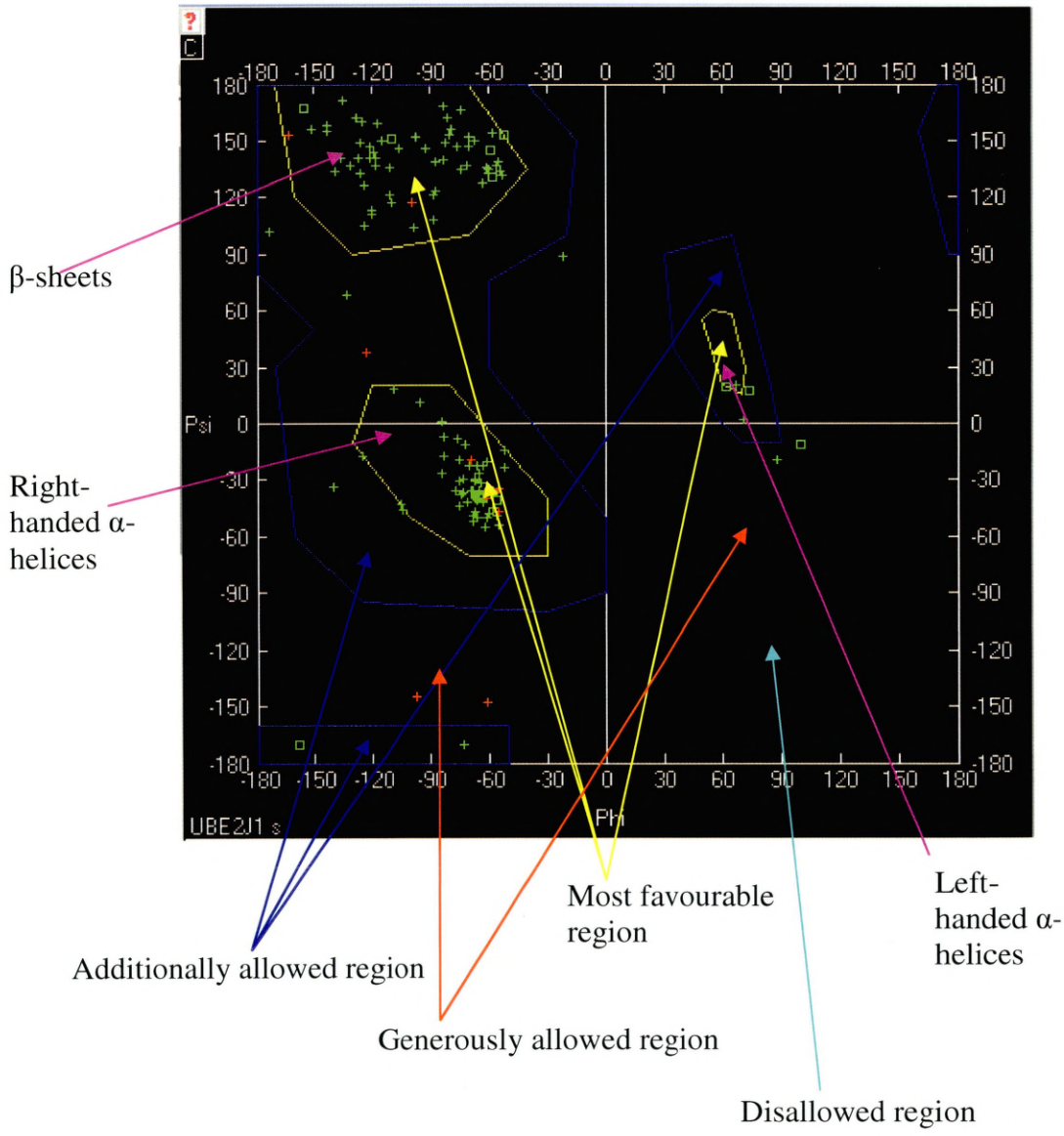
Figure 15.6 is the model of UBE2J1 coloured by B-factor which is a measure of how much structural information the modelling program was able to transfer from the template. As could be seen from the figure above, that the model complies with the criteria, where it could be said that it has homology along the whole length of the peptide sequence, except for the region in red where the 9 amino acids are missing in the template and hence there is no homology in that region.

The next step of evaluation was to check the Ramachandran plot, the details of which and all other details of the What Check report are illustrated as follows:

15.1.3.2. The Ramachandran plot

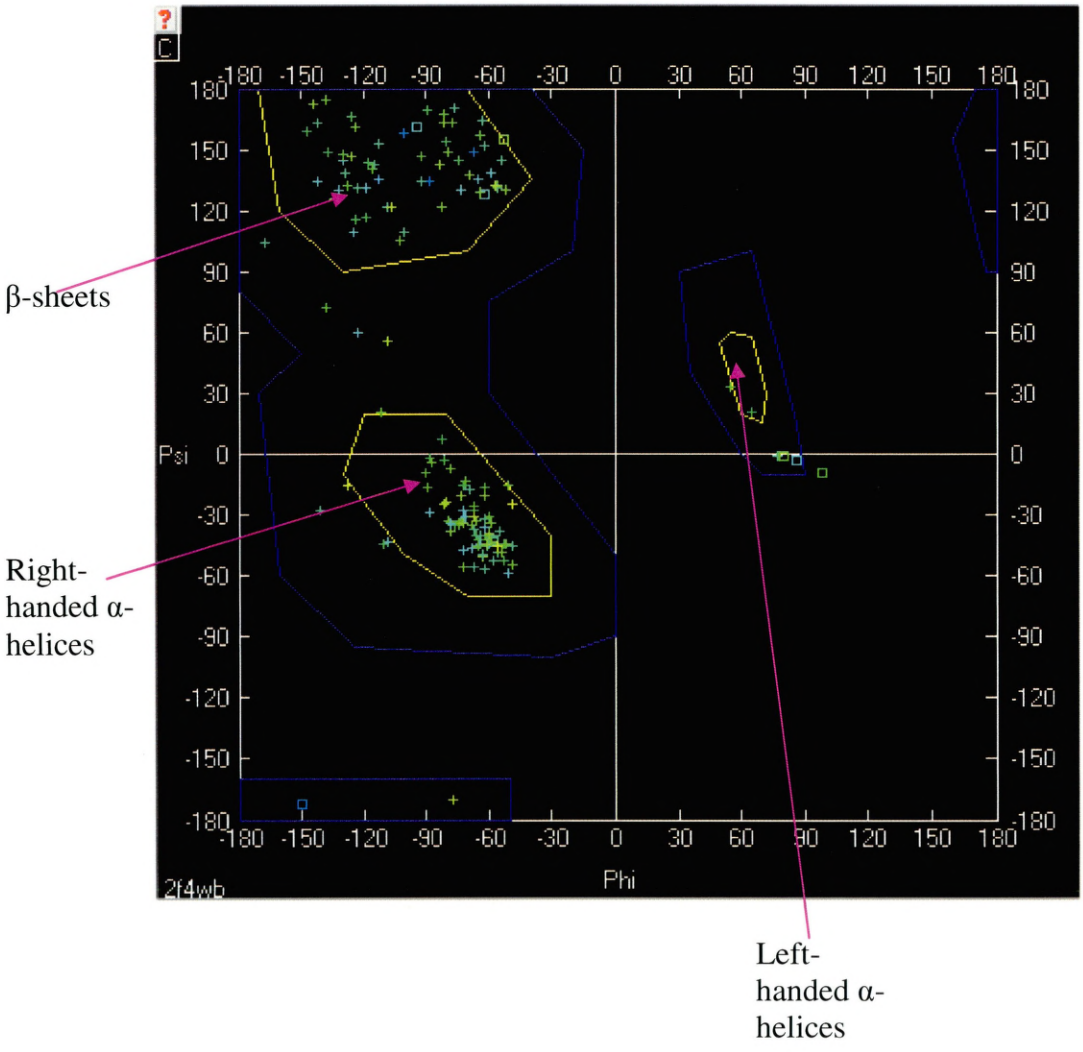
The most powerful and important tool to check the stereochemical quality of a protein structure, is the Ramachandran Plot (Ramachandran, Ramakrishnan and Sasisekharan, 1963). The plot is of a Phi versus Psi (main chain torsion angle) for every amino acid residue of the protein with the exception of two terminal residues, because the N-terminal residue does not have Phi and the C-terminal residue has no Psi. The resulting scatter plot is a cluster of points (of amino aids) in certain favourable regions, and some points tend to be excluded into the disallowed regions due to steric hindrances of the side chain atoms. Glycine is an exception as it has no side chains, and therefore can adopt phi and psi angles in all four quadrants of the Ramachandran plot. Hence it frequently occurs in turn regions of proteins where any other residue would be sterically hindered, showing slightly different distributions on the plot. The Ramachandran plots of the model UBE2J1 (Figure 15.7.1), and the template 2F4WB (Figure 15.7.2) are as follows:

Figure 15.7.1 Ramachandran plot of modelled UBE2J1



Figures 15.7.1 shows the Ramachandran plot of the model of UBE2J1 protein. The small plus indicates any amino acid other than glycine and the square sign indicating glycine in the different regions. There is the favourable region where regular secondary structures like helices, beta sheets and even residues in loops also lie in this region, and there is also the disallowed region where amino acids are located which is due to steric hindrances or clashes between atoms, with the exception of glycine.

Figure 15.7.2 The Ramachandran plot of the template (2F4WB)



Figures 15.7.2 shows the Ramachandran plot of the template 2F4WB. UBE2J2 has most amino acids in the" most favourable" region.

It can be seen from the Ramachandran plots of both the template UBE2J2 and the model UBE2J1 that most amino acids are in the "most favourable" region; very few amino acids are in the "additionally allowed" regions. There are two amino acids of the UBE2J1 model (GLU91 & THR92) that are in the generously allowed region.

15.1.3.3. Force field energy

The next step was to calculate the average energy for each group within a structure, based on its confirmation and interaction with its neighbours. This gave a list of the energy states of each amino acid in the model, shown in Figure 15.8.

Figure 15.8 Force field energy of UBE2J1 model

RESIDUES	RES. NO.		FORCE FIELD ENERGY	RESIDUE	RES. NO.		FORCE FIELD ENERGY
PRO	1	E=	-17.198	ASP	43	E=	20.097
ALA	2	E=	-19.268	PHE	44	E=	39.791
VAL	3	E=	-2.627	ASP	45	E=	26.476
LYSH	4	E=	-17.644	GLY	46	E=	68.074
ARG	5	E=	-221.073	GLY	47	E=	10.378
LEU	6	E=	-28.644	VAL	48	E=	-21.374
MET	7	E=	-36.208	TYR	49	E=	66.534
LYSH	8	E=	1.653	HISA	50	E=	-19.339
GLU	9	E=	-21.272	GLY	51	E=	-1.521
ALA	10	E=	-21.741	ARG	52	E=	-244.666
ALA	11	E=	-27.932	ILE	53	E=	-8.810
GLU	12	E=	-0.372	VAL	54	E=	-24.652
LEU	13	E=	-20.606	LEU	55	E=	36.636
LYSH	14	E=	-5.675	PRO	56	E=	-3.421
ASP	15	E=	120.714	PRO	57	E=	-4.978
PRO	16	E=	-0.440	GLU	58	E=	-17.588
THR	17	E=	5.559	TYR	59	E=	-26.173
ASP	18	E=	0.197	PRO	60	E=	-28.091
HISA	19	E=	-28.501	MET	61	E=	6.050
TYR	20	E=	47.852	LYSH	62	E=	-7.121
HISA	21	E=	-19.472	PRO	63	E=	5.316
ALA	22	E=	-21.770	PRO	64	E=	-28.052
GLN	23	E=	-166.574	SER	65	E=	-27.675
PRO	24	E=	-27.051	ILE	66	E=	35.991
LEU	25	E=	-27.751	ILE	67	E=	89.223
GLU	26	E=	5.574	LEU	68	E=	17.130
ASP	27	E=	16.196	LEU	69	E=	-26.907
ASN	28	E=	-200.650	THR	70	E=	-22.627
LEU	29	E=	20.971	ALA	71	E=	9.577
PHE	30	E=	-19.934	ASN	72	E=	-99.180
GLU	31	E=	-41.806	GLY	73	E=	31.416
TRP	32	E=	-32.248	ARG	74	E=	-276.236
HISA	33	E=	-56.421	PHE	75	E=	34.272
PHE	34	E=	-52.709	GLU	76	E=	-17.460
THR	35	E=	-37.963	VAL	77	E=	38.154
VAL	36	E=	-12.361	GLY	78	E=	48.077
ARG	37	E=	-247.172	LYSH	79	E=	21.875
GLY	38	E=	36.025	LYSH	80	E=	-28.933
PRO	39	E=	-4.606	ILE	81	E=	26.377
PRO	40	E=	-2.541	CYSH	82	E=	-4.041
ASP	41	E=	0.976	LEU	83	E=	72.207
SER	42	E=	-28.019	SER	84	E=	-9.086

Force field energy of UBE2J1 model continued to the next page

Figure 15.8 Force field energy of UBE2J1 model continued:

RESIDUE	RES. NO.		FORCE FIELD ENERGY
ILE	85	E=	26.523
SER	86	E=	4.349
GLY	87	E=	36.139
HISA	88	E=	31.156
HISA	89	E=	30.609
PRO	90	E=	9.666
GLU	91	E=	18.901
THR	92	E=	102.200
TRP	93	E=	142.671
GLN	94	E=	-150.682
PRO	95	E=	16.796
SER	96	E=	-0.143
TRP	97	E=	31.435
SER	98	E=	6.329
ILE	99	E=	25.426
ARG	100	E=	-162.579
THR	101	E=	-17.243
ALA	102	E=	-30.387
LEU	103	E=	-12.578
LEU	104	E=	-5.792
ALA	105	E=	-32.999
ILE	106	E=	32.973
ILE	107	E=	197.139
GLY	108	E=	30.154
PHE	109	E=	-1.069
MET	110	E=	2.240
PRO	111	E=	56.796
THR	112	E=	-1.191
LYSH	113	E=	44.301
GLY	114	E=	44.236

RESIDUE	RES. NO.		FORCE FIELD ENERGY
GLU	115	E=	40.704
GLY	116	E=	33.527
ALA	117	E=	34.640
ILE	118	E=	162.800
GLY	119	E=	18.950
SER	120	E=	-32.325
LEU	121	E=	-2.577
ASP	122	E=	-7.976
TYR	123	E=	-67.092
THR	124	E=	8.203
PRO	125	E=	-0.290
GLU	126	E=	-14.879
GLU	127	E=	-48.248
ARG	128	E=	-277.930
ARG	129	E=	-270.661
ALA	130	E=	-17.043
LEU	131	E=	91.393
ALA	132	E=	-19.757
LYSH	133	E=	-36.385
LYSH	134	E=	147.443
SER	135	E=	-28.719
GLN	136	E=	-183.476
ASP	137	E=	-24.073
PHE	138	E=	-24.127
CYSH	139	E=	-18.677
CYSH	140	E=	-12.490
GLU	141	E=	56.150
OXT	141	E=	11.336
/----------			----------
KJ/mol		E=	-1463.023

Figure 15.8 shows the list of energy states of all amino acids of the model UBE2J1 generated by DeepView. The lower the value, the more stable the interactions among residues, where large positive values indicate unstable interactions among residues. Very few amino acids have high force field energy like THR92, TRP93, ILE107, ILE118 and LYS134. Most other amino acids have force field energy within acceptable limits.

Here it can be seen that the overall energy state of the model UBE2J1 is low (around -1463 KJ/mol) and hence the model is stable.

The different force fields are the GROMOS, AMBER, CHARMM, and SYBYL. "GROMOS" was developed by Van Gunsteren. GROMOS was used in DeepView to calculate force fields, which can calculate force fields, perform energy minimization and molecular dynamics simulation (constant energy or temperature, constant volume or pressure), in vacuo, in crystals or in aqueous or any other solutions.

The equation for GROMOS force field calculations is as follows:

$$
\begin{aligned}
&V(\vec{r}) \\
&= \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 \\
&+ \sum_{\text{dihedrals}} K_\chi(1+\cos(n\chi-\delta)) \\
&+ \sum_{\text{nonbonded-pairs}\,i,j} \left[\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} - \varepsilon_{ij} \left\{ \left(\frac{R_{\min ij}}{r_{ij}} \right)^{12} - 2\left(\frac{R_{\min ij}}{r_{ij}} \right)^{6} \right\} \right]
\end{aligned}
$$

The above equation was adapted from (MacKerell [No date] http://www.wiley.com/legacy/wileychi/ecc/samples/sample07.pdf)

This equation could be separated into the internal terms, which would include the bond, angle, dihedral contributions, and the nonbonded or external terms that include the electrostatic and van der Waals terms. b_0 and θ_0 represents the equilibrium bond and angle terms, K_b, K_θ, and K_x represents the bonds, angles, and dihedral force constants, n represents the multiplicity or periodicity of the dihedral term, Δ represents the phase of the dihedral term, and b, θ, and x are the bond, angle, and the dihedral angle respectively that define **r**. The dihedral term is a fourier series, indicating an accurate description of torsional surfaces. q_i and q_j are the partial atomic charges on atoms i and j. e_0 is the dielectric constant, r_{ij} is the distance between atom i and j. ε_{ij} and $R_{min\ i\ j}$ are the well depth and minimum interaction distance describing the van der Waals interaction between atoms i and j. (Clark, 1997;

http://amrit.ittc.ku.edu/tclark/bio97/node1.html; Gunsteren, 2001;

http://www.igc.ethz.ch/gromos-docs/index.html).

The different force fields are as follows:

AMBER- (Assisted Model Building and Energy Refinement)- widely used for proteins and DNA.

CHARMM- originally developed at Harvard, widely used for both micro molecules and macro molecules.

GROMOS- A force field that comes as part of the GROMOS (GROningen MOlecular Simulation package) a general-purpose molecular dynamics computer simulation package for the study of biomolecular systems. GROMOS force field (A-version) has been developed for application to aqueous or apolar solutions of proteins, nucleotides and sugars. However, a gas phase (B-version) for simulation of isolated molecules is also available.

SYBYL- This force field was developed for the calculation of internal geometries and conformational energies.

(Cornell et al., 1995;

http://www.molvis.indiana.edu/app_guide/InsightII/forcefields.html)

GROMOS, CHARMm force fields are the category of simple diagonal force fields. These force fields incorporate a simple harmonic diagonal representation for the bond and angle terms. GROMOS and CHARMm incorporate the Van Der Waals parameters derived from crystal data, whereas the Van Der Waals parameters in the AMBER force field are derived from liquid simulations. The AMBER and GROMOS force fields specify values for A (repulsive coefficients), and B (attractive coefficients), whereas CHARMm specify values for R* and ε. CHARMm uses arithmetic mean combining rules for R*, and geometric mean combining rules for ε. By using different values for A and B for a particular atom type, GROMOS makes a further distinction, depending on the second atom involved in the interaction. While all force fields incorporate a simple fourier expansion to represent the dihedral energy, some variations are also seen in the assignment of the energy. GROMOS reduces the NBF computational workload by

excluding interactions beyond a (user input) cut off radius, an important fact in parallel algorithm selection. A pair list indicating which non-bonded interactions are computed every t steps, where t is typically in the range from 10 to 50. The data structures representing the pair list tend to be the most space consuming in the program, with hundreds of interaction partners for each atom (Cornell et al., 1995).

It has been seen from the force field energy of UBE2J1, that there are no such high values which could create a problem in the structure.

15.1.3.4. Colour by force field energy

The energy parameters of a protein can predict many features of interaction within it. Potentials of mean force derived from the statistical analysis of interaction regularities in proteins, can recognise misfolded structures or assess the quality of models prepared by homology modelling and capture changes to protein structure model during refinement (Godzik, 2003).

As it takes a long time to go through the energy states of each residue, it is easier to colour the model by force field energy to see the energy states of the residues. So the next step was to colour the structure by force field energy, where in general long stretches of warm colours are indications that the model is not stable. The structure of UBE2J1 as coloured by force field energy is as shown in Figure 15.9.

Figure 15.9 Model of UBE2J1 coloured by force field energy

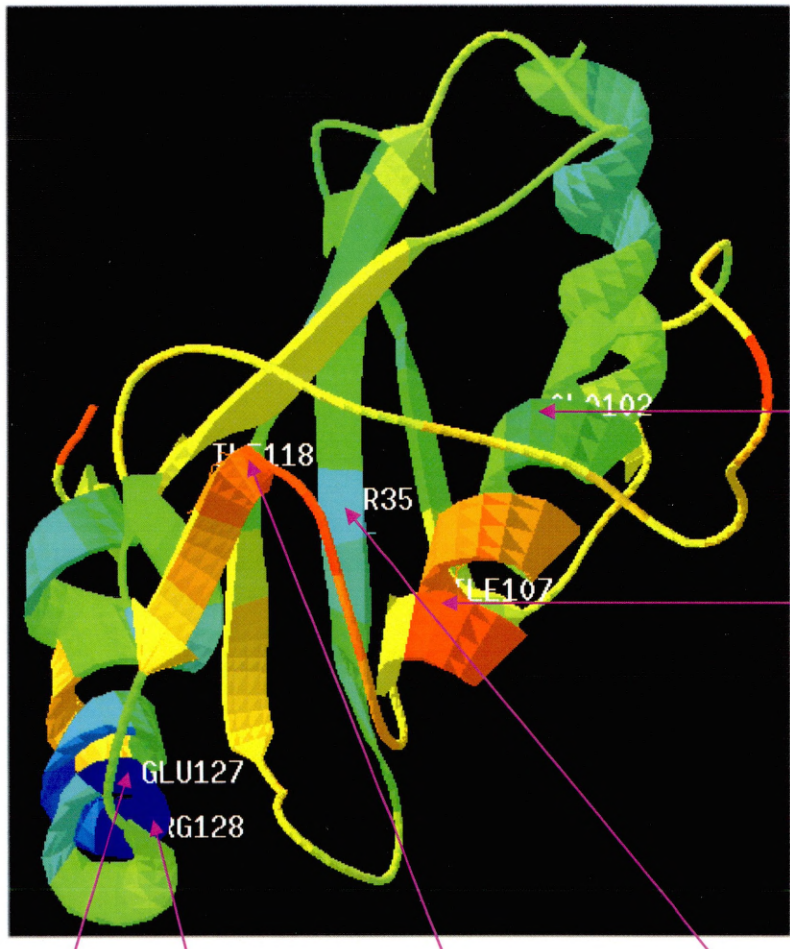


Figure 15.9 shows the model of UBE2J1 coloured by force field energy. Amino acids with different force field energies have been illustrated in the figure. Amino acids like ILE107 and ILE118 has higher force field values of 197.1 and 162.8 KJ/mol respectively, whereas amino acids like ALA102 has force field energy of -37.9 KJ/mol, THR35 -37.9 KJ/mol, ARG128 -277 KJ/mol and GLU127 -48.2 KJ/mol has force field values in the negative range indicating contributing to the stability of the protein. Here it can be seen that most of the regions of the model is having low force field energies, clearly stating that the model generated is stable.

From the Figure 15.9 it can be seen that although some of the regions in the model shows instability, overall the model is stable as can be seen from the Figure 15.9.

Hence from all the above checks it was found that the model of UBE2J1 is stable which is further confirmed by the following results illustrated as follows:

15.1.3.5. Model of UBE2J1 coloured by RMS (root mean square)

Figure 15.10 Colour by RMS

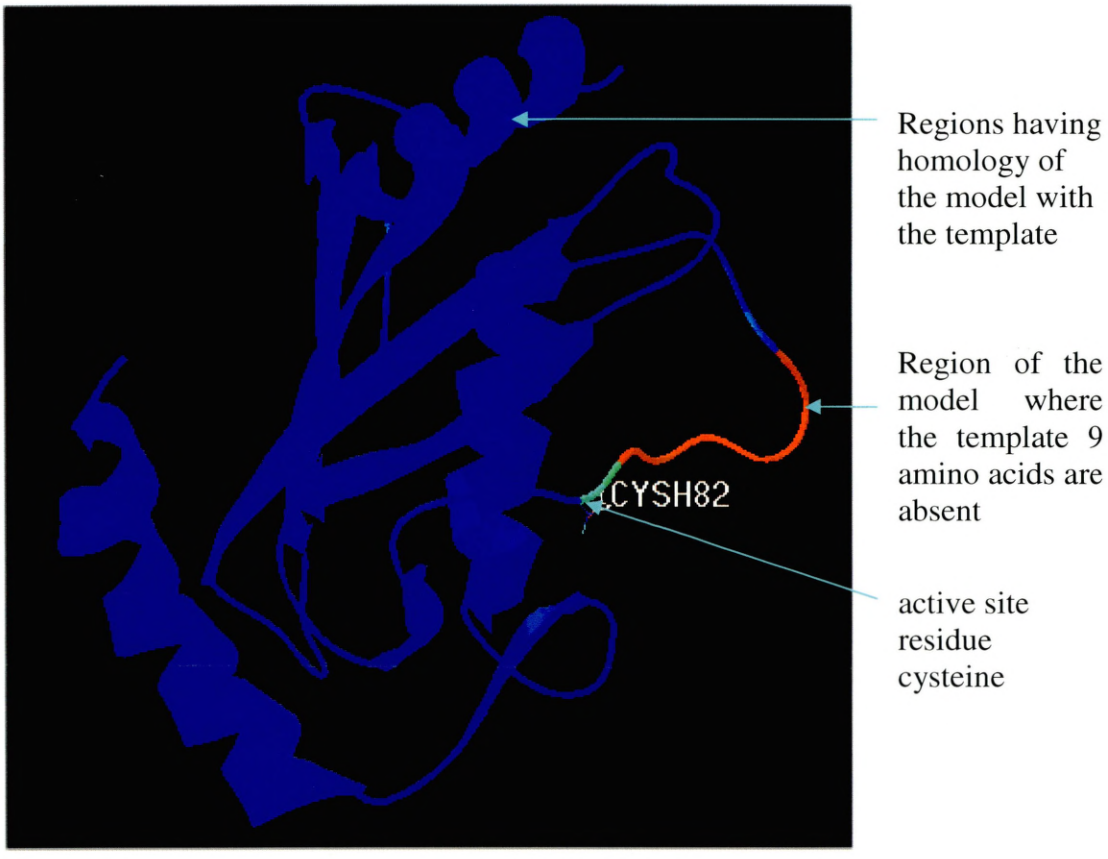


Figure 15.10 shows the root mean square deviations (RMSD) of the model UBE2J1 from the template (2F4WB) structure. The only region that shows the model to be divergent from the template is the region marked in red, as the 9 amino acids in that region of the template structure are not available.

The model UBE2J1 and the template structure (2F4WB) were compared to find regions which are similar or divergent, by colouring by root mean square (RMS). As can be seen from Figure 15.10, except for the region of the missing 9 amino acids in the template which shows divergence, most other regions of the model and template have similarity. The deviation of the model from the template is further illustrated by alignment diversity as in figure 15.11.

15.1.3.6. Model of UBE2J1 coloured by alignment diversity

Figure 15.11.1 Colour by alignment diversity

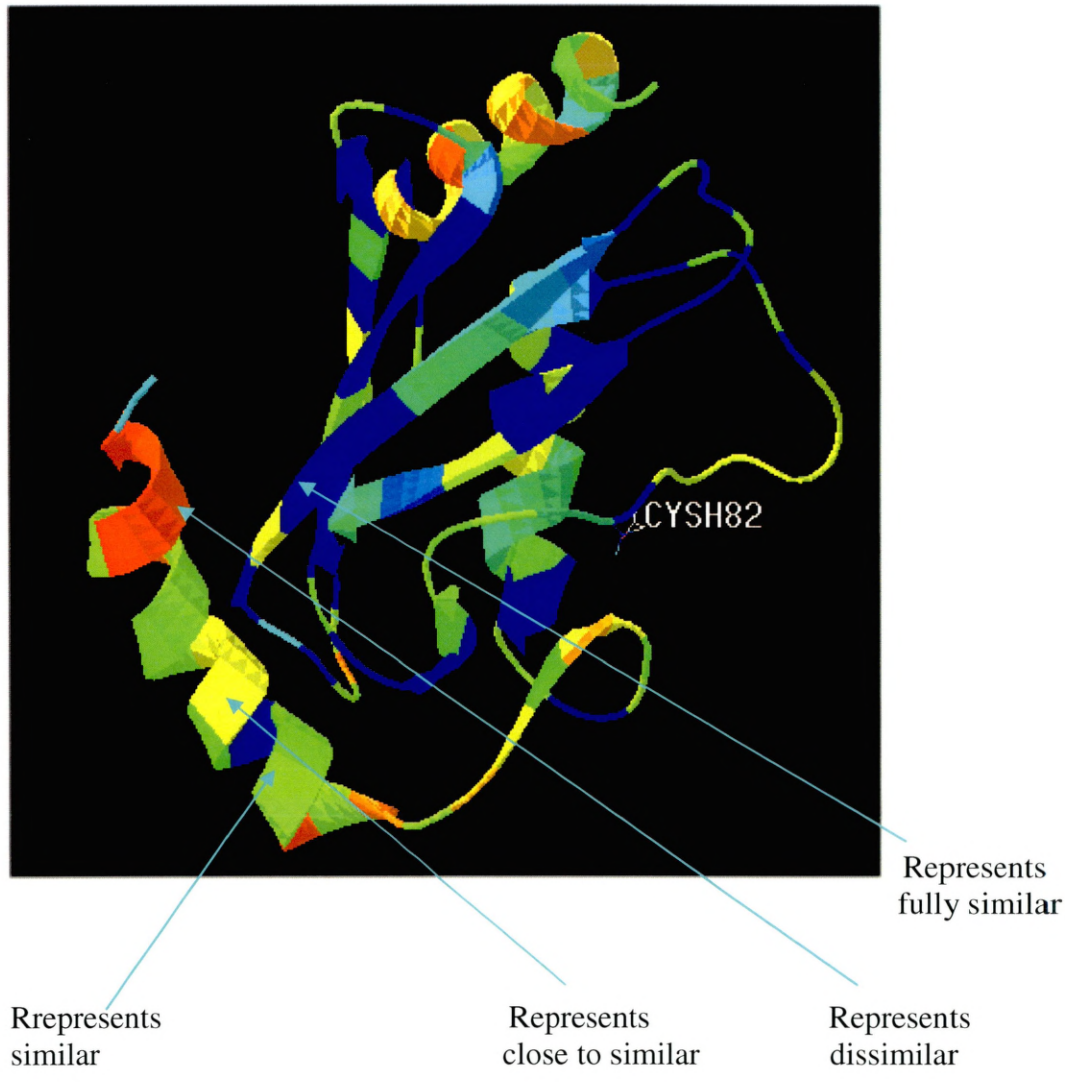


Figure 15.11.1 illustrates the model of UBE2J1 coloured by alignment diversity, of the level of similarity of UBE2J1 with the template structure 2F4WB. It can be seen from the model coloured by alignment diversity that, most of the regions of the model has similarity with the template, except for very short regions which shows divergence between the model and the template.

The alignment diversity is further illustrated by the sequence alignment of the model and the template as in Figure 15.11.2.

Figure 15.11.2 Colour by alignment diversity shown in the alignment

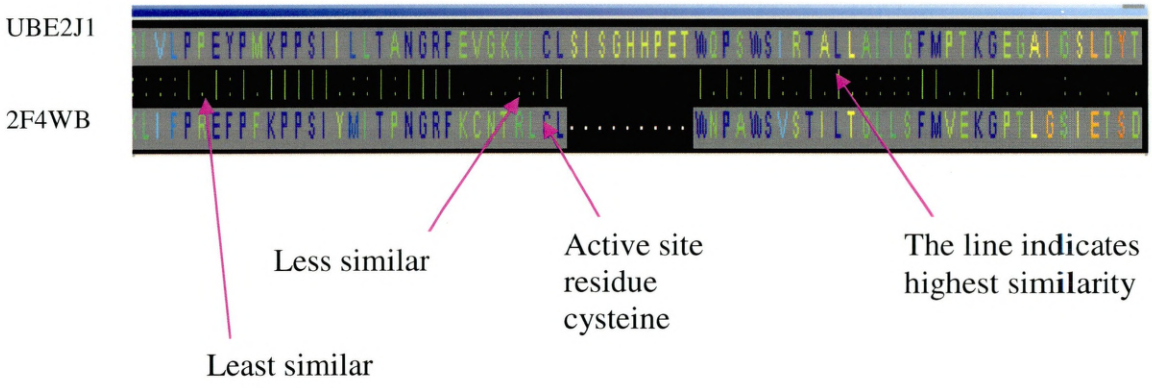


Figure 15.11.2 shows the same colouring by alignment diversity of the model of Figure 15.11.1, but at the sequence alignment level. As can be seen the line between the amino acids coloured in blue indicates the highest similarity, with a double dot between two amino acids coloured in green indicating similarity but not as much as amino acids coloured in blue. This is followed by the single dot between amino acids coloured in yellow indicating even less similarity.

When a model and its template are compared and their peptide sequences been aligned, the option of colour by alignment diversity helps to colour the sequences by the degree of conservation in their primary sequence. The different levels of conservation as can be seen in the alignment Figure 15.11.2, indicates that the amino acids are completely conserved, conserved, close to conserved and not at all conserved.

15.1.3.7. Model of UBE2J1 coloured by secondary structure

Figure 15.12 Colour by secondary structure

Figure 15.12 shows the model of UBE2J1 coloured by secondary structure. Here it shows all the β sheets coloured in green, α helices coloured in yellowish brown and the loops in grey.

15.1.3.8. Model of UBE2J1 coloured by solvent accessibility

Figure 15.13.1
Colour by solvent accessibility of the UBE2J1 model

Figure 15.13.2
Colour by solvent accessibility of the UBE2J2(2F4W) structure

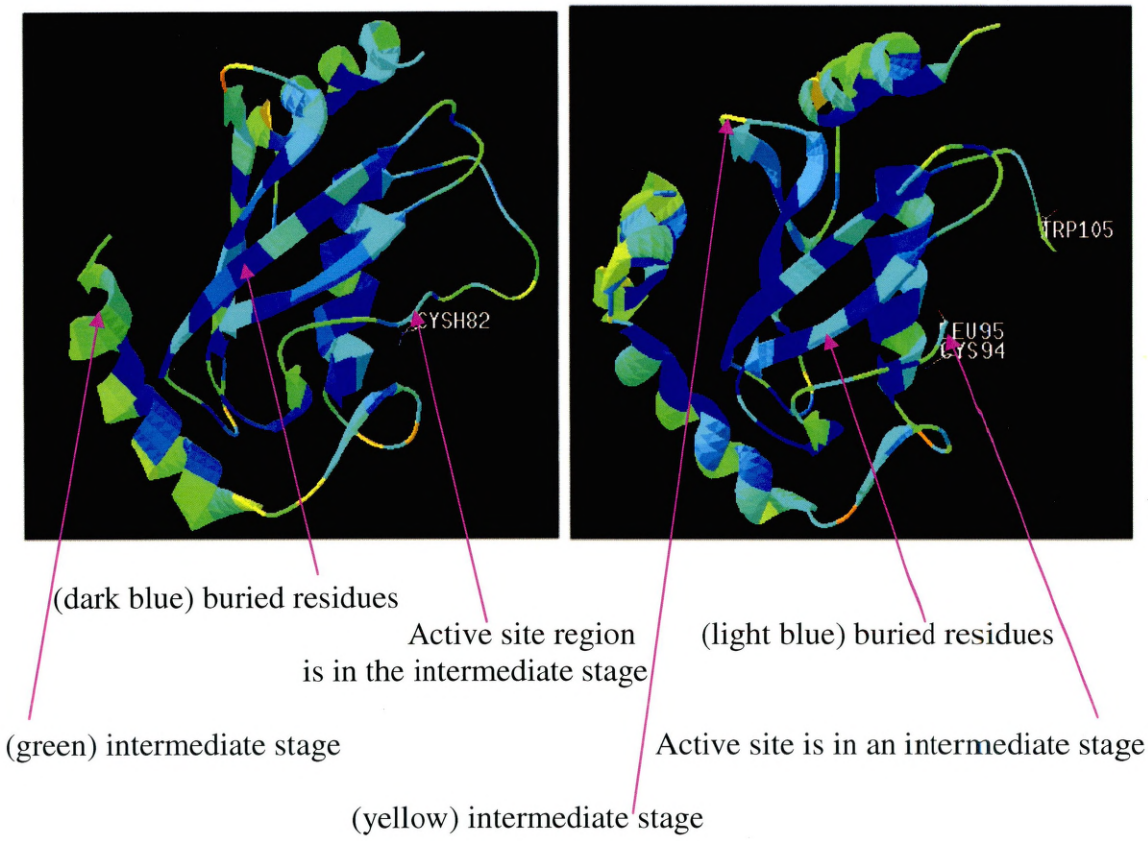


Figure 15.13.1 shows the model of UBE2J1, and Figure-15.13.2 shows the template (2F4WB), coloured by residues according to their solvent accessibilities. The more exposed amino acids to the solvent are in red (not present in either UBE2J1 model or in the UBE2J2 structure), the intermediately exposed amino acids in green and yellow, and the buried residues in blue. Here there are no amino acids of both UBE2J1 and the template UBE2J2 that are exposed to solvent.

Solvent accessibility can be used as a criterion towards 3D structure prediction, the principle goal of which is to predict the extent to which a residue embedded in a protein structure is accessible to solvent (Lee and Richards, 1971; Chotia, 1976; Connolly, 1983). The volume of a residue embedded in a structure that is exposed to solvent, where different residues has different accessible surface areas. Residues that are buried have a relative solvent accessibility of less than 16%, whereas those that are exposed have solvent accessibility 16%.

The accessibility option will colour the residues according to their solvent accessibilities. This is calculated as a percentage of theoretical surfaces of the amino acid which is accessible to water. Buried residues are coloured blue (not accessible) whereas more exposed amino acids become green and then red.

From the Figures 15.13.1 and 15.13.2, of the UBE2J1 model and the template 2F4WB, it can be seen that both the model and the template, have their active sites CYS exposed to the solvent, as it is appears green in colour. Although an amino acid should appear red in colour so as to show its complete exposure to the surface, still the green colour of the active site CYS indicates that although it is not completely exposed, but still it tends towards the surface.

15.1.3.9. What check report comparison of the template (2F4WB) and the model of UBE2J1

Table 15.1 Whatcheck report of the template 2F4WB and the UBE2J1 model

Structure Z-scores, positive is better than average:		
	Template (2F4WB)	model of UBE2J1
1st generation packing quality	: -0.225	-1.658
2nd generation packing quality	: 0.176	-2.071
Ramachandran plot appearance	: -0.890	-0.321
chi-1/chi-2 rotamer normality	: -0.515	2.026
Backbone conformation	: 0.912	0.159
RMS Z-scores, should be close to 1.0:		
Bond lengths	: 0.742	0.780
Bond angles	: 0.803	1.468
Omega angle restraints	: 1.072	0.991
Side chain planarity	: 1.005	2.250 (loose)
Improper dihedral distribution	: 0.964	1.972 (loose)
B-factor distribution	: 1.008	---
Inside/Outside distribution	: 1.073	1.170 (unusual)

Table 15.1 is a comparative study of the Whatcheck reports of the template UBE2J2 (pbd ID 2F4WB) and the model of UBE2J1 generate by DeepView. The values of the above table are the Z-scores or the number of standard deviations away from the mean. The terms mentioned "loose" and "unusual" against the values indicated that they deviate from the standard value.

The measured value minus the mean value, divided by the standard deviation is the Z-score which is represented by the equation as below:

$$Z = X - mu/ sigma.$$

If the value of "X" is less than the mean then the Z-score value is negative, but is positive if the value of "X" is greater than mean. The Z-score value should be around 0, and when it is positive, it is greater than the average, but when it is negative it is worse than the average.

The RMS-Z score is calculated as per the equation as given below:

RMS-Z = $\sqrt{}$ sum (Z^2)/ number of bonds.

The RMS-Z score value should be around 1.0.

15.1.3.10. Feature of the Whatcheck report:

Packing quality: In the analysis of protein packing quality, the threading potential is made use of, which expresses how well the peptide sequence sits in the structure. Whatif calculates whether a configuration is favourable or not, by making use of the database densities. The occurrence of all possible atoms in all possible positions around the fixed fragment (any largest group of atoms that does not contain a torsion angle, is a fixed fragment) is counted.

Ramachandran plot: Also known as the Phi/Psi plot, is the first verification tool for protein structures. The Ramachandran plot has the alpha helix and the beta strand area and also some loop conformations area. The Ramachandran plot allows very few residues to lie outside these areas as mentioned. The exception is glycine, as it does not have a side chain, and hence has not restriction.

chi-1/chi-2 rotamer: As all amino acids have their different preferences for the chi1/chi2 rotamers, Whatif calculates the chi1/chi2 conformation Z-scores.

Backbone conformation: Whatif checks for the number of times a similar backbone conformation exists in its database. If the number of times the conformation exists is less than 3, then the amino acid residue is said to have a unique backbone conformation. The structural average score, that describes how well the hits in the backbone database, fits the structure, referred to as the backbone normality of the query structure.

Bond lengths: As bond lengths are very well known entities, there are standards, set in proteins which includes average and standard deviations for all different bond types. Using these standards, the Z-scores for each of the bonds in the protein structure are

calculated by Whatif. All Z-scores of ($Z < -4$ or $Z > 4$) are listed in the Whatcheck report.

Bond angles: Similar to the bond lengths, the average and standard deviations of bond angles are obtained and used to calculate the Z-scores for all bond angles.

Omega angle restraints: In theory omega angles are restrained to 180 degrees, with a small amount of flexibility for small molecular structures. Variation in the omega angles is determined by Whatif, and noted in the report.

Side chain planarity: Residues that have large planarity deviations could cause strain in the structure, which could be caused by incorrect backbone or side chain positioning. All residues that deviate more than 4 times the normal value from planarity are listed.

Improper dihedral distribution: The measure of the chirality or planarity of the structure at a specific atom is called Improper dihedral. 0 degree will be obtained if the planar atoms, and for non-planar configurations (chiral atoms), values are around +35 and -35 degrees. The mean value and standard deviation for the corresponding Improper dihedral, for each of the atom types in the protein, is calculated by Whatif from its database of 500 X-ray structures.

Inside/Outside distribution: As all hydrophobic residues like to sit on the "inside" and the hydrophilic residues on the "outside" of the globular protein, Whatif calculates an overall RMS Z-score to show how all residues are located in protein. Ideally for globular proteins the RMS Z-score is close to 1.

Listed above is an overall comparison of the Whatcheck report, of the template structure (UBE2J2), with that of the model of UBE2J1 obtained by homology modelling in DeepView. From the whatcheck reports it can be said that the structure of UBE2J2, has all its geometric parameters within permissible limits. The model of UBE2J1 has some of the geometric parameters whose values are within permissible limits, like the Ramachandran plot, the bond length, bond angles, and the omega angle restraints. There are few geometric parameters of the model of UBE2J1, whose values are not within the permissible limits. For example, the side chain planarity, and the improper

dihedral distribution values of the model of UBE2J1 are termed as "loose" by the whatcheck report. Also, the Inside/Outside distribution value is also termed as "unusual" by the whatcheck report whose normal values are between 0.84 and 1.16. The second generation packing quality value is also -2.929, which is not suitable whatcheck value criteria. Generally, a second generation packing quality value below -3.0 is considered a poor model and a value of below -5.0 is definitely a wrong model.

15.1.4. Homology modelling by 3D-JIGSAW

To compare the prediction of the 3D model of UBE2J1 initially obtained by DeepView, another online homology modelling had been tried to predict the model of UBE2J1 by an online server called 3D-JIGSAW, results of which is as shown and discussed as follows:

The UBE2J1 peptide sequence had been modelled by 3D-JIGSAW, by using UBE2J2 (2F4WB) as the template. The following is the model of UBE2J1 modelled by 3D-JIGSAW (Figure 15.14).

Figure 15.14

Model of UBE2J1 obtained by homology modelling from 3D-JIGSAW, coloured By secondary structure

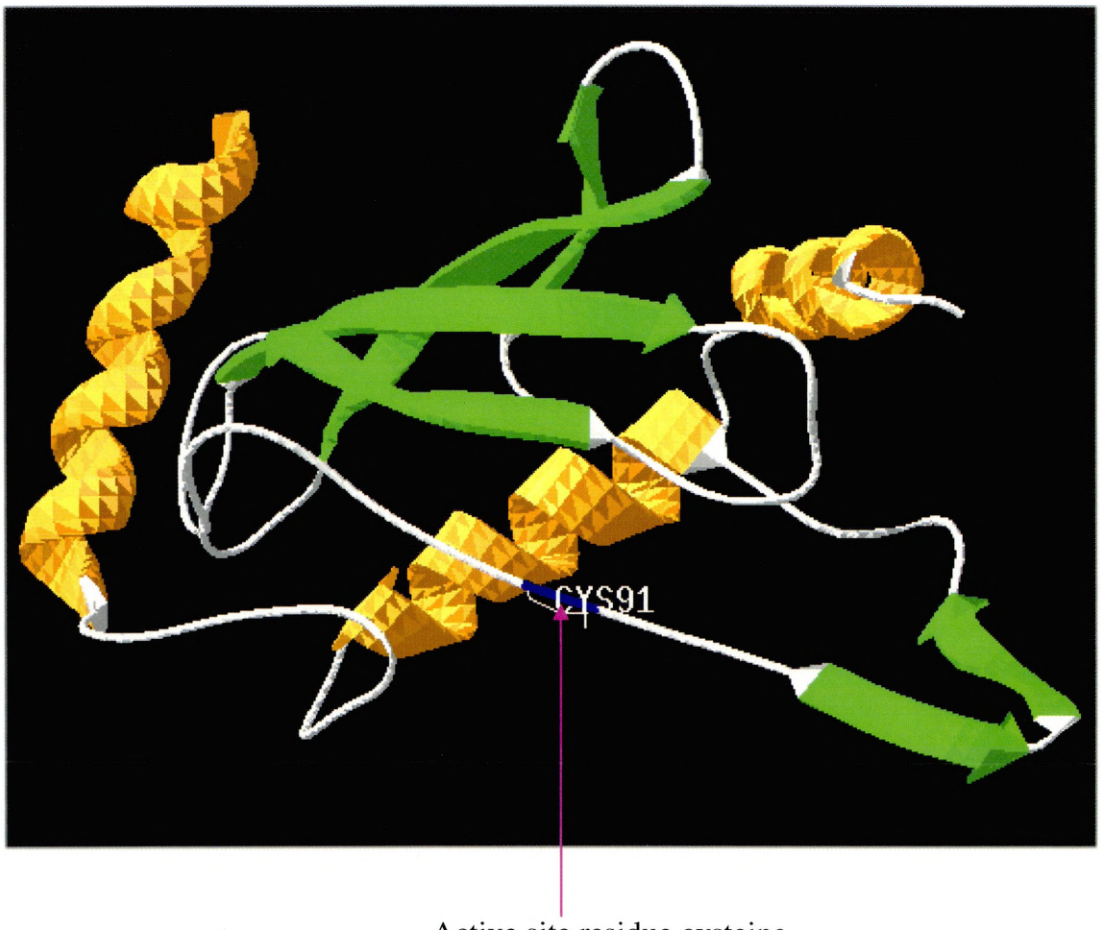


Figure 15.14 is the model of UBE2J1, obtained by homology modelling from 3D-JIGSAW online homology modelling server. This model is coloured by B-factor indicating a high degree of conservation of amino acids of the model generated, with that of the template structure (2F4WB).

In order to see the possible conformations of the Phi and Psi angles of amino acid residues in the modelled UBE2J1 structure, the Ramachandran plot of the model of UBE2J1 obtained from 3D-JIGSAW is shown in Figure 15.15.

Figure 15.15

The Ramachamdran plot of the model of UBE2J1, obtained from 3D-JIGSAW homology modelling

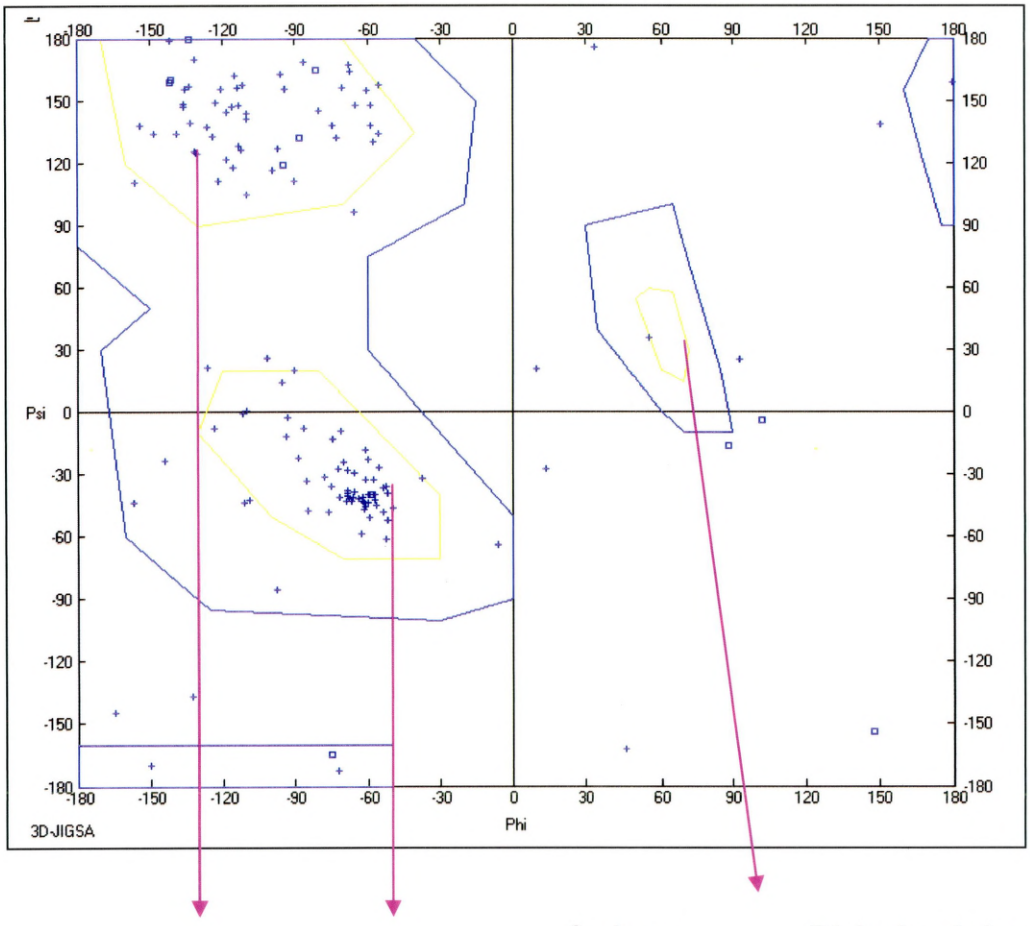


Figure 15.15 shows the Ramachandran plot of the model of UBE2J1 obtained by homology modelling in 3D-JIGSAW online homology modelling server. The small plus indicates any amino acid other than glycine and the square sign indicating glycine in the different regions.

From the Ramachandran plot shown in Figure 15.15 of the model of UBE2J1 generated by 3D-JIGSAW, it can be seen that most amino acids lie in the most favourable region, very few lie in the additionally allowed and the generously allowed regions.

15.1.4.1. What check report comparison of the model of UBE2J1 obtained by homology modelling by DeepView, with that of the UBE2J1 model obtained from 3D-JIGSAW.

Each reported fact has an assigned severity, one of:

* **error**: severe errors encountered during the analyses. Items marked as errors are considered severe problems requiring immediate attention.
* **warning**: Either less severe problems or uncommon structural features. These still need special attention.
* **note:** Statistical values, plots, or other verbose results of tests and analyses that have been performed. If no error was found, this will also be listed as a note.

1) Bond angle variability

Model of UBE2J1 by DeepView

Note: Normal bond angle variability. Bond angles were found to deviate normally from the mean standard bond angles. The RMS Z-score given below is expected to be around 1.0 for a normally restrained data set, and this is indeed observed for very high resolution X-ray structures. More common values are around 1.55

RMS Z-score for bond angles: 1.468

RMS-deviation in bond angles: 2.849

Model of UBE2J1 by 3D-JIGSAW

Warning: High bond angle deviations

Bond angles were found to deviate more than normal from the mean standard bond angles. The RMS Z-score given below is expected to be around 1.0 for a normally restrained data set, and this is indeed observed for very high resolution X-ray structures. More common values are around 1.55. The fact that it is higher than 2.0 in this structure might indicate that the constraints used in the refinement were not strong enough. This will also occur if a different bond angle dictionary is used.

RMS Z-score for bond angles: 2.116

RMS-deviation in bond angles: 4.194.

Interpretation: The model obtained by DeepView has normal bond angle variability, whereas the model obtained from 3D-JIGSAW have high bond angle variability.

2) Ramachandran Z-score

Model of UBE2J1 by DeepView

Note: Ramachandran Z-score OK
The score expressing how well the backbone conformations of all residues are corresponding to the known allowed areas in the Ramachandran plot is within expected ranges for well-refined structures.

Ramachandran Z-score : -0.321

3D-JIGSAW model
Warning: Ramachandran Z-score low

The score expressing how well the backbone conformations of all residues are corresponding to the known allowed areas in the Ramachandran plot is a bit low.

Ramachandran Z-score : -3.040

Interpretation: The overall impression of the Ramachandran plot is the Ramachandran Z-score, where the negative value of the Z-score is worse than the average and positive is better than the average. As the negative value gets lower than
-3, Whatif starts giving an indication that it is low.
The model obtained from Deepview has the Ramachandran Z-score just below the average, but the model obtained from 3D-JIGSAW has low Z-score value.

Table 15.2 gives a comparison of Whatcheck reports of the model of UBE2J1 generated by Deepview and that generated by 3D-JIGSAW.

3) Overall comparison of the Whatcheck reports of the models of UBE2J1 obtained by Deepview and 3D-JIGSAW

Table 15.2 Whatcheck report comparison of the models of UBE2J1 generated by Deepview and 3D-JIGSAW

Structure Z-scores, positive is better than average:		
	model of DeepView	3D-JIGSAW model
1st generation packing quality	: -1.658	-2.434
2nd generation packing quality	: -2.071	-3.259 (poor)
Ramachandran plot appearance	: -0.321	-3.040 (poor)
chi-1/chi-2 rotamer normality	: 2.026	-1.817
Backbone conformation	: 0.159	-2.538
RMS Z-scores, should be close to 1.0:		
Bond lengths	: 0.780	2.237 (loose)
Bond angles	: 1.468	2.116 (loose)
Omega angle restraints	: 0.991	2.427 (loose)
Side chain planarity	: 2.250 (loose)	8.141 (loose)
Improper dihedral distribution	: 1.972 (loose)	3.600 (loose)
Inside/Outside distribution	: 1.170 (unusual)	1.120

Table 15.2 is a comparative study of the Whatcheck reports of the model of UBE2J1 and the model generate by 3D-JIGSAW. The terms indicated as "poor", "loose" and "unusual" along with some of the values are the values which deviate very much from the required value. The Z-score should be around 0 and the RMS-Z score should be around 1.0.

15.1.4.2. Interpretation of the Whatcheck report of the model generated by DeepView and the model generated by 3D-JIGSAW :

Whatcheck criteria had been used to check the model of UBE2J1, generated by 1) deepview, and 2) 3D-JIGSAW, it can be said that the model of UBE2J1 generated by Deepview, is better than the model generated by 3D-JIGSAW. The value of second generation packing quality of the UBE2J1 model generated by 3D-JIGSAW is represented as "poor" by whatcheck. The Ramachandran plot appearance of the UBE2J1 model generated by 3D-JIGSAW is also represented as "poor". The bond lengths, bond angles, and omega angle restraints values of the UBE2J1 model generated by 3D-JIGSAW, are all represented as "loose" by the whatcheck report. Hence, from the above mentioned quality checks it can clearly be said, that the model of UBE2J1 generated by Deepview, is better than that generated by 3D-JIGSAW.

Although the model generated by DeepView complies with almost all criteria of a good model, this model will not be ideal for predicting the substrate binding site, at the region where there is no homology around 9 amino acids as illustrated in Figure 15.6. The substrate binding site in this model could however be predicted at other regions of the active site, especially the regions where other amino acid residues like G, R, F which are unique in the UBC6.

15.2. Structural comparisons of UBE2Js (UBE2J1 & UBE2J2) with UBC9 protein

The following section gives the results of the structural superimposition of UBE2J1 and UBE2J2 with UBC9.

The two human orthologues of yeast UBC6 are the UBE2J1 and UBE2J2. As Avvakumov et al. (2006) predicted the structure of one of the human homologues of yeast UBC6, UBE2J2 (pdb ID- 2F4W) from the co-ordinates obtained from X-ray crystallography. The UBE2J2 structure had been compared to yeast UBC9 (pdb ID 1U9A) so as to see how the UBC6 family PROSITE signature (active site region) fit in to that of UBC9. The PROSITE signature obtained for the UBC6 family is quite different from that of all other UBC.

The reason for choosing UBC9 as an example for the structural comparisons mentioned before is as follows:

More information on the active site, of UBC9 has been known than any other UBCs. Again from the prosite signatures of UBC6 and non-UBC6 UBCs it is quite evident that cysteine in the PROSITE signature is the most important catalytic residue. As previously implicated by Pickart (Pickart, 2001) and recently shown by Yunus & Lima, and Knipscheer & Sixma (Yunus and Lima, 2006; Knipscheer and Sixma, 2006), that Asparagine ASN85 of UBC9 is also important in the catalysis. In addition to Asparagine, Tyrosine TYR87 and aspartic acid ASP127 of UBC9 has also been said to have catalytic activity. This is illustrated in Figure 15.16.

Prosite signature of all UBCs except UBC6: [FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x-[LIV].*

Figure 15.16 The catalytically active amino acids in UBC9

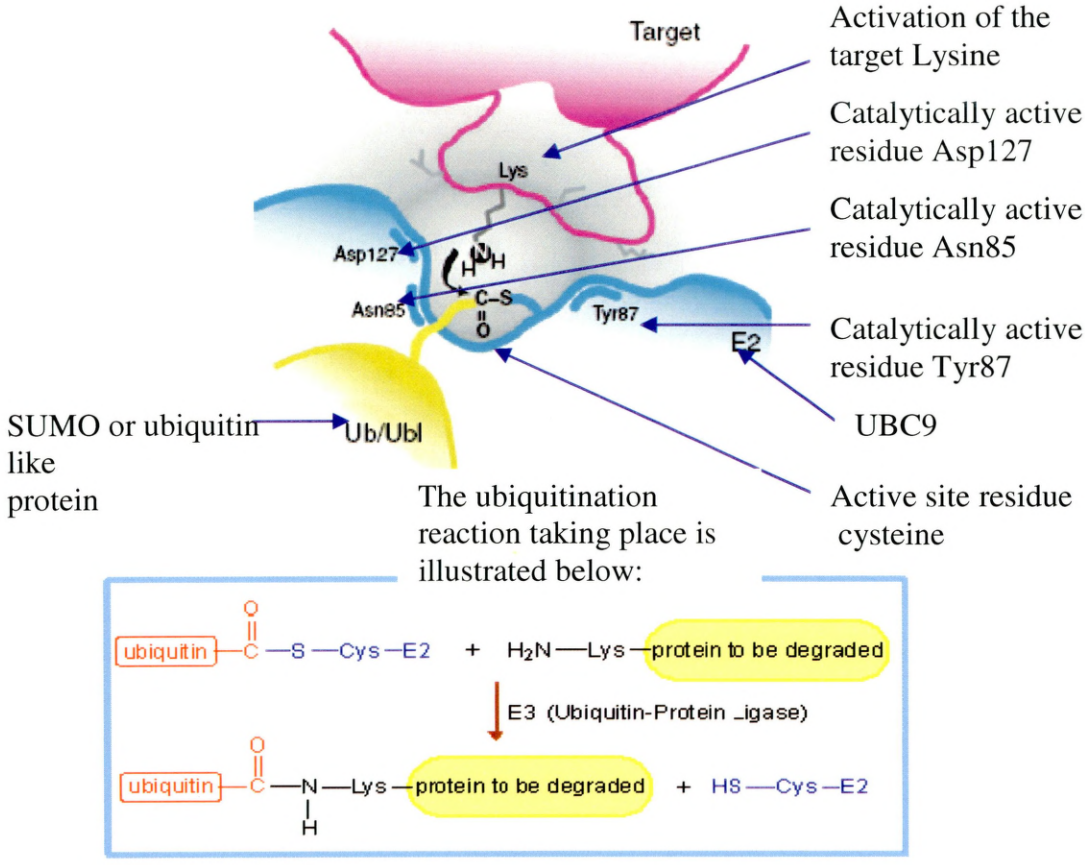


Figure 15.16 shows three residues in the UBC9 active site, Asn85, Tyr87 and Asp127 having the catalytic activity, other than the active site residue cysteine. These three residues (Asn85, Tyr87 and Asp127) in addition to the catalytic cysteine are located in a pocket that helps to position the target (p53) lysine (the only titratable residue). Illustrated in the figure is the reaction between the target protein and the E2-SUMO thioester complex. The figure shows the activation of the lysine residue of the target protein, with a decrease in its pK value by Asn85, Tyr87 and Asp127 (figure was adapted from Knipscheer and Sixma, 2006). The ubiquitination reaction taking place illustrated in the above figure was adapted from (http://www.rpi.edu/dept/bcbp/molbiochem/MBWeb/mb2/part1/protease.htm).

No clear picture of the E2 mechanism has been identified so far. Asparagine residue (Asn85) was identified by Pickart and coworkers but no catalytic acid or base had been found so far. Recently Yunus and Lima has effectively used a screening experiment in yeast, in which the human UBC9 complements the essential yeast UBC9 (Yunus and Lima, 2006). A number of UBC9 residues had shown to have an effect on the growth rate, but none of them, in contrast to the catalytic cysteine is essential for yeast survival. Asn85, Tyr87, and Asp127, has shown to affect the rate of conjugation reaction, but not affect the affinity for the substrate. Cysteine together with these three residues are located in a pocket from where it helps in the interation with the lysine, but any mutation of the three amino acids Asn85, Tyr87, and Asp127 causes the destabilization of the lysine side chain conformation. (Knipscheer and Sixma, 2006). Several kinetic studies were carried out where it showed that the decrease or increase in the pK of the target lysine was affected by these there amino acids Asn85, Tyr87, and Asp127 along with the active site residue cysteine.

15.2.1. Superimposition of UBE2J2 with UBC9

Figure 15.17

Superimposition of UBE2J2 with UBC9

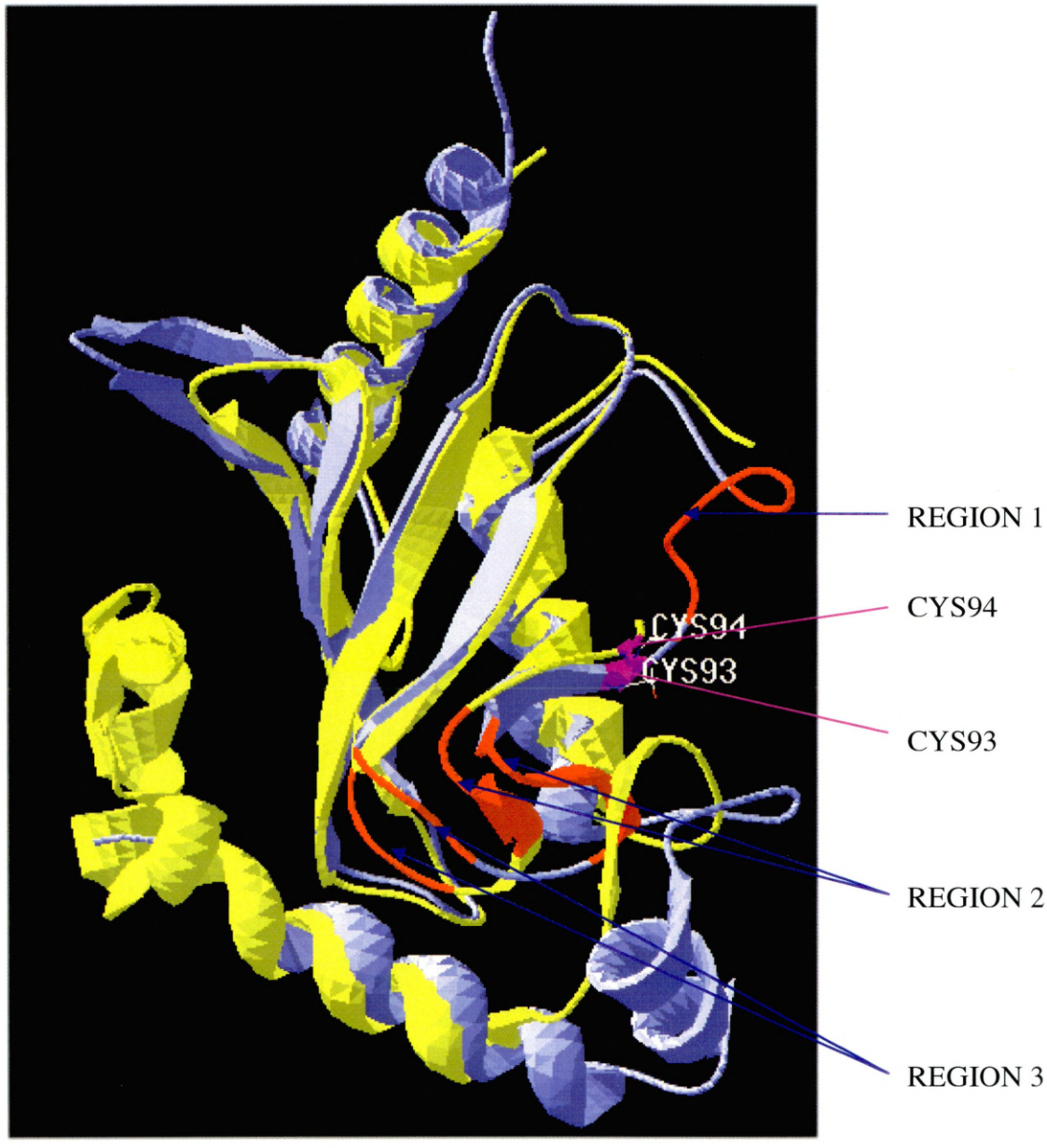


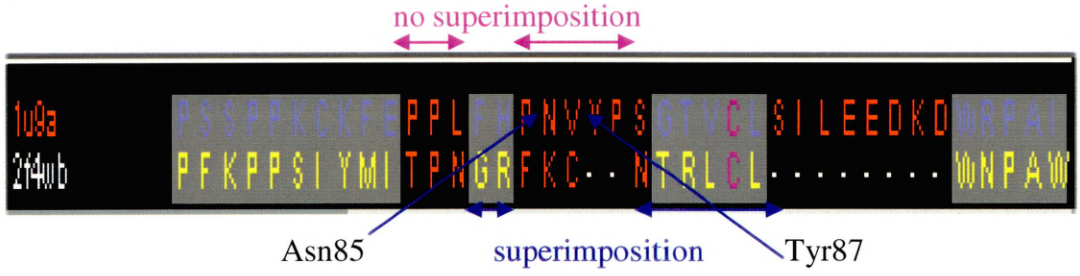
Figure 15.17 illustrates the superimposition of UBE2J2 (coloured in yellow) with UBC9 (coloured in light blue). Illustrated in the figure are the CYS94 (of UBE2J2) and CYS93 (of UBC9) coloured in pink. Three non-superimposing regions nearest to the active site residue cysteine of UBE2J2 and UBC9 are marked in red and marked in the figure as REGION 1, REGION 2 and REGION 3. The RMS of the superimposition of UBE2J1 to UBC9 is 1.15Å.

Figure 15.18.1

Structural superimposition of UBE2J2 and UBC9 of Figure 15.17, shown at its amino acid residue level.

Non-UBC6 UBCs PROSITE signature:

[FYWLPS]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C*-[LIV]-x-[LIV]



Proposed UBC6 PROSITE signature:

T-[PAR]-[NS]-G-R-F-x(3)-[KTE]-[RK]-[LIV]-C*-[LMS]-[ST]-[IMF]-[ST]-x(2)-H-[PK]

Figure 15.18.2

Structural superimposition of UBE2J2 and UBC9 shown at its amino acid residue level, continued from the previous Figure 15.18.1.

Figure 15.18.1 shows a short region at the amino acid level (around the active site) of the superimposition of the UBC9 (pdb ID- 1U9A) structure with the UBE2J2 (pdb ID- 2F4WB) structure (illustrated in Figure 15.17). The reason to show the structural superimposition at the amino acid residue level is to illustrate better which amino acids do not superimpose as it was not possible to show all of them in the structure in Figure 15.17. The PROSITE signatures of both UBC6 and Non-UBC6 UBCs PROSITE signature has also been shown.

Also shown in Figure 15.18.1 are the catalytically active amino acids of UBC9 which are asparagine (Asn85) and tyrosine (Tyr87). The third catalytically active amino acid of UBC9 which is Aspartic acid (Asp127) is shown in Figure 15.18.2. From both the Figures 15.18.1 & 15.18.2, it can be seen that these catalytically active residues of

UBC9 do not superimpose with UBE2J2, hence indicating that these residues are catalytically important for UBC9 only and not to UBE2J2 nor any other UBCs.

It can be seen from the structural superimposition of UBE2J2 with UBC9 that surprisingly the active site residue cysteine of both the structures superimposes. This was an unexpected result as UBE2J2 and UBC9 have very different substrate specificities. Therefore other amino acids surrounding the active site cysteines must be responsible for changing the substrate specificity. As can be seen from figure 15.17 and 15.18, there appears to be three areas of non-superimposing amino acids near the active sites of UBE2J2 and UBC9 that may be responsible for the different substrate specificities. The first and nearest area to the active site cysteine of UBE2J2 at amino acid 94, is the sequence from amino acids 87 to 90, namely F, K, C, N, where F is the invariant phenylalanine in our proposed UBC6 PROSITE signature. The second region that does not superimpose on the UBC9 structure, near the active sites are the UBE2J2 amino acids from 82 to 84 namely T, P, N, where T is the invariant threonine of the proposed UBC6 PROSITE signature. To the other side of the active site cysteine (towards the C-terminal) as there are 9 amino acids missing in the 2F4WB structure, there is this region of non-superimposition with the structure of UBC9. This is highlighted in Figure 15.17 as "REGION 1".

Also, as the catalytically active residues Asn85, Tyr87 and Asp127 of UBC9 do not superimpose with UBE2J2, it could be hypothesised that these three residues are only catalytically important for UBC9, but may not have any substrate specificity for any other UBCs.

15.2.2. Superimposition of UBE2J1 and UBC9

Figure 15.19 Superimposition of UBE2J1 and UBC9

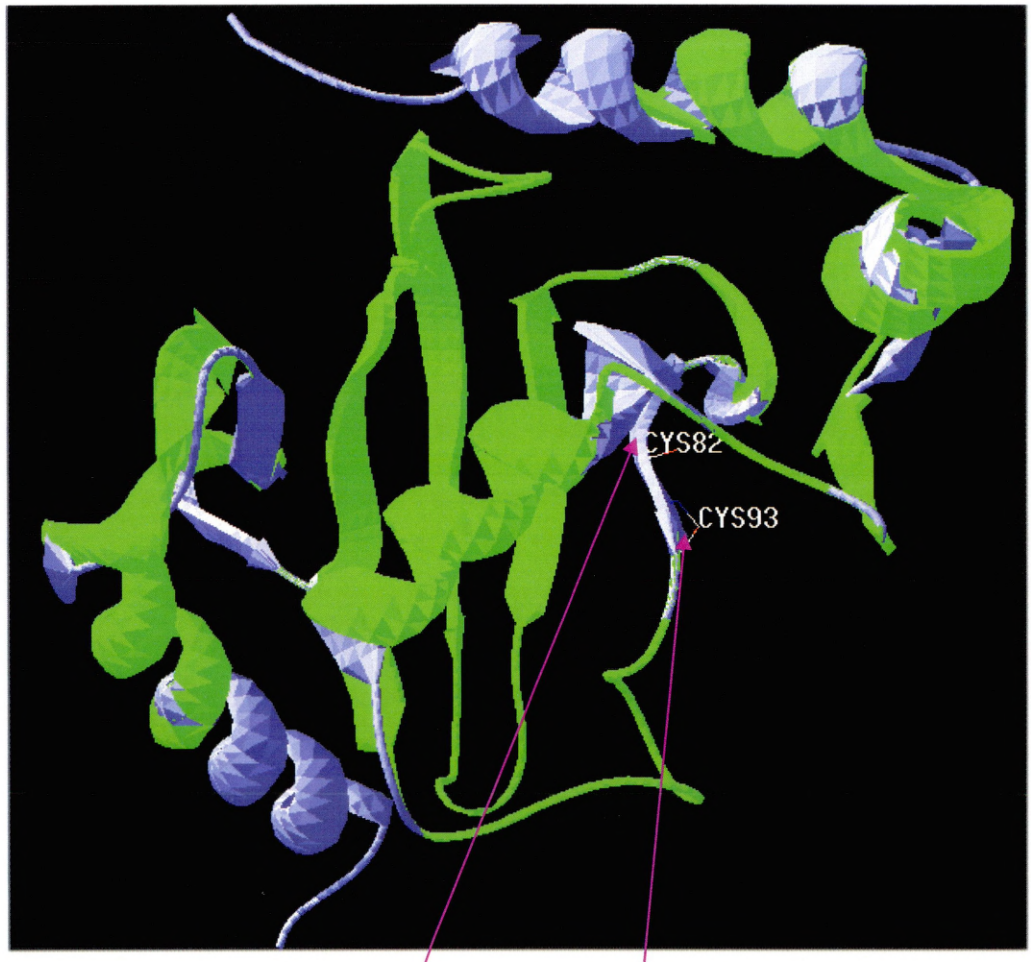


Figure 15.19 shows the 3D structure of human UBC9 (pdb ID 1U9A) coloured in light blue, consists of 158 amino acids have been superimposed with the model of UBE2J1 coloured in green consisting of 141 amino acids.

From the superimposition of UBE2J1 with UBC9 it can be seen that both superimposes completely. The conservation of the amino acids of UBE2J1 and UBC9 is illustrated in Figure 15.20.

Figure 15.20

Conservation of UBE2J1 and UBC9 along the whole length of the peptide sequence

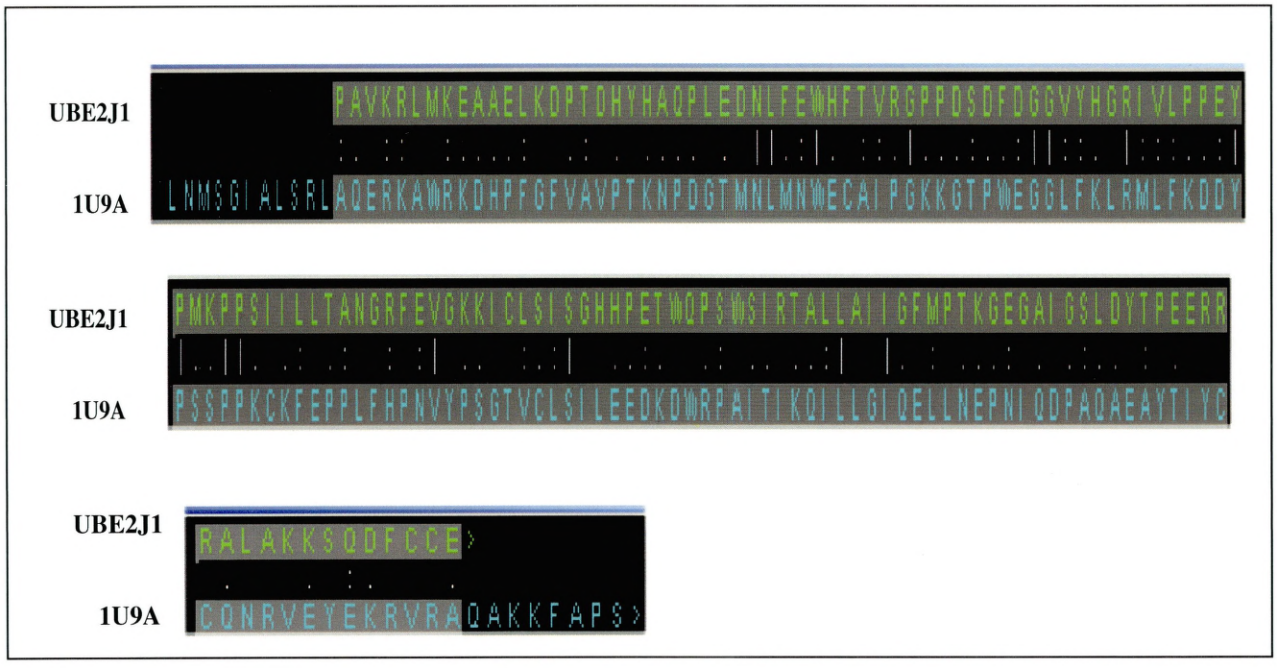


Figure 15.20 illustrates the conservation of UBE2J1 with UBC9 (pdb ID 1U9A) along the whole length of its peptide sequences.

As can be seen from Figure 15.19, both the UBE2J1 model and UBC9 structure is completely superimposed. This is in contrast to the structural superimposition of UBE2J2 with UBC9 where there are various regions of non-superimposition, especially along the active site region as illustrated in Figure 15.17 & 15.18. These differences in the structural superimpositions of UBE2J1 with UBC9 and UBE2J2 with UBC9 states clearly that the DeepView generated model of UBE2J1 is unique and its structural findings will have a strong impact for further studies, except the region where there is no homology with the template (2F4WB), where the 9 amino acids are missing.

15.3. Conclusion

The model obtained by homology modelling of UBE2J1 is in a good agreement with the practical approach of homology modelling. Using the approach described above, the model of UBE2J1 was built based on the coordinates of 2F4WB structure. The Z-score & RMS Z-scores obtained from analysis by WHATIF (Whatcheck report), showed that the model obtained is acceptable.

The best available template that had been used here which had a low percentage identity (around 43%) but within acceptable limits (as theoretically the percentage identity should be 30% and above). By doing the alignment corrections, and various other optimization steps, the model obtained had been checked and verified and had been found to have satisfied most of the evaluation criteria.

The superimpositions of UBE2J2 with UBC9 had shown that the UBC6 family differs from all other UBCs, especially with respect to the active site and have different substrate specificities. There were no diffraction patterns obtained for the 9 amino acids of the crystal of UBE2J2 (2F4WB) by Avvakumov et al. (2006), as these 9 amino acids are highly mobile or disordered. It could be hypothesised from this that this mobility and disorder of amino acid of UBE2J2 could be the same as in UBE2J1. This could be also one of the reasons why the UBE2J1 failed to crystallize.

15.4. Future research

The future research would be to carry out a docking analysis, to find out a suitable substrate that would specifically bind to the active site of human UBE2Js (UBE2J1 and UBE2J2), and not to any other UBCs. This would be required to inhibit the degradation function of UBE2Js and would hopefully have therapeutic implications in the treatment of cystic fibrosis.

CHAPTER SIXTEEN

OVERALL CONCLUSION AND FUTURE RESEARCH

16.1 Overall conclusion and future research

Sequences of homologous proteins from related organisms are very similar. The difference grows with an increase in the evolutionary distance between species. The observation has led to the development of sequence alignment methods (Doolittle, 1996) where the optimum alignment can be found between the sequences of two or more proteins being compared. Following the sequence alignment, the phylogenetic analysis is carried out of those aligned sequences, so as to draw a evolutionary and functional relationship of the sequences and the species from which the sequences were derived. By carrying out the phylogenetic analysis of all the 13 yeast UBCs and their respective homologues from eleven selected species, an evolutionary and functional relationship was inferred. Another important use of the phylogenetic tree was the proposal for the correct nomenclature of all wrongly named UBCs. The HGNC nomenclature system needs to assign unique names to proteins, on the basis of well established phylogenetic relationships. We have successfully renamed the NCUBE1 and NCUBE2 as UBE2J1 and UBE2J2 respectively, and would propose a system of nomenclature for all UBC (E2s) on the basis of the interpretations of the phylogenetic tree of the 193 peptide sequences.

Results of the Consurf analysis has produced a phylogenetic tree of the same peptide sequences used for phylip in the studies. This phylogenetic study has investigated the structural relationship of UBE2J2, to all its homologues in UBC (E2s). The plot represents different levels of conservations (Figure 12.3) and provides a better understanding of the conservation of UBC (E2s).

The identification of the previously unidentified UBC (E2) active site of the *Drosophila* $TAF_{II}250$ has opened up a new area of further research. Moreover the identification of the active site residue Cysteine in only the *Apis mellifera* has led to the assumption of convergent/divergent evolution where again future research could be carried out in the laboratory to confirm the finding of the UBC (E2) active site region in *Drosophila* $TAF_{II}250$.

Proteins are described and classified into families and analysed for patterns of mutations at various positions along the sequence. The sequence similarity between proteins from

different species forms the basis of molecular phylogenetics. Protein structures have a complex shape defined by specific interactions between amino acids along the chain. Different sequences adopt similar folds. For closely related proteins, it is possible to build reliable evolutionary trees and analyze the relationships between the organisms from which they came. For studying the dynamics of side chain movements in proteins or for predicting the results of a point mutation, it may not be thought of as a necessity of how enzymes evolved. But there are problems that require insight from both perspectives. Why does proteins performing different functions have very similar structures and the vis-a versa (Godzik, 2003)? The homology modelling, protein fold assessment and ab initio protein structure prediction are the very important bioinformatics tools, which brings mechanistic insights to biology, chemistry and medicine. A significant evolutionary change in the gene sequence is easily noticed at the level of an individual protein functional unit or domain. Structural biologists can contribute enormously to protein structure, of which homology modelling can be distinguished from all other methods for analysing relationships among protein sequences, because it produces atomic co-ordinates suitable for direct comparisons with X-ray and NMR structures (Burley and Bonanno, 2003).

We have tried to predict the tertiary structure of the protein UBE2J1 by X-ray crystallography. This approach has proved so far, to be unsuccessful because of the failure to grow usable protein crystals. We have used an alternative strategy to generate a tertiary structure for UBE2J1 using computational structure prediction (homology modelling). UBE2J1 and UBE2J2 are the two human homologues of yeast UBC6, which has a sequence identity of around 43%. Although the crystal structure of UBE2J1 could have given a more confident prediction, the structure obtained by homology modelling could give valuable insights into the structural similarity of UBE2J1 with UBE2J2. The future research could be to design primers of the UBE2J1 protein (as mentioned in sections 14.3 and 14.4), by taking other suitable regions of the UBE2J1 protein, as predicted by Globplot and RONN (globular or disordered region prediction program). By carrying out various trials, could possibly lead to crystal growth in at least one of the UBE2J1 designed primers.

16.2 The use of Grid to identify specific UBE2J inhibitors

As the X-ray crystallographic structure of UBE2J2 is now known (Avvakumov et al., 2006), another area of future research would be to find a substrate that would bind specifically to the yeast UBC6 homologues, i.e. human UBE2J1 and UBE2J2, and not bind to any other UBCs of all the 13 yeast UBC family. Such inhibitors may prove to be of therapeutic value in the treatment of cystic fibrosis. The inhibition of ERAD using a UBE2J1/J2 inhibitor should stop the degradation of ΔF508 CFTR protein, thus making functional ΔF508 CFTR proteins available for use into the plasma membrane. The design of suitable inhibitors can be helped by using the very recently available GRID system (http://www.grid-support.ac.uk/) comprising programmes such as "AutoDock (Morris et al., 1998; 1996; Goodsell and Olson, 1990) or flexX" (Lengauer and Klebe, 1996).

Autodock is a software that runs under the grid agent which tries to determine how small molecules, such as substrates or drugs, can bind to a receptor of a known 3D structure. There are three separate programs in Autodoc: The docking of the ligand to a set of grids describing the target protein is performed by Autodoc; These grids are pre-calculated by AutoGrid; and finally the AutoTors program sets up which bonds will be treated as rotatable in the ligand

(http://www.scripps.edu/mb/olson/doc/autodock/; http://www.grid-support.ac.uk/).

One of the first success's achieved by AutoDock was called WISDOM (Wide In Silico Docking On Malaria http://wisdom.eu-egee.fr/malaria/). This system enables the relatively rapid identification of compounds that bind specifically to 3D protein structures using "ZINC", which is a free database of commercially available compounds (Irwin and Shoichet, 2005).

The Zinc database contains over 3.3 million compounds available for docking experiments. Hence from the database the compound that best fits to the receptor of the 3D structure is found, with the help of computers running the analysis at various nodes in various parts of the world, hence increasing the CPU capacity that is required for such a high capacity of space for the analysis.

Our structural superimposition of human UBE2J2 on UBC9 (UBE2I) shows that using GRID programmes such as Autodoc, with the ZINC database, it would be possible to identify molecules that bind specifically to the active sites of UBE2J's but not to other UBCs. Such molecules as mentioned earlier may serve as therapeutic agents to cystic fibrosis patients.

CHAPTER SEVENTEEN

REFERENCES

Adams, J. 2002. Proteasome inhibition: a novel approach to cancer therapy. *Trends Mol Med.* 8(4): pp.S49-S54.

Adams, P.D., Grosse-Kunstleve, R.W. and Brunger, A.T. 2003. Computational aspects of high-throughput crystallographic macromolecular structure determination. In: Bourne, P.E. and Weissig, H. eds. *Structural Bioinformatics*. New Jersey: Wiley. pp.75-87.

Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219: pp.555-565.

Altschul, S.F., Madden, T.L., Schaffer, A.A, Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389-3402.

Amerik, A.Y. and Hochstrasser, M. 2004. Mechanism and function of deubiquitinating enzymes.*Biochim Biophys Acta.* 1695(1-3): pp.189-207.

Andersson, D., Carlsson, U. and Freskgard, P.O. 2001. Contribution of tryptophan residues to the CD spectrum of the extracellular domain of human tissue factor: application in folding studies and prediction of secondary structure. *Eur J Biochem.* 268(4): pp.1118-1128.

Andersen, P.L. ,Zhou., H., Pastushok, L., Moraes, T., McKenna, S., Ziola, B., Ellison, M.J., Dixit, V.M. and Xiao, W. 2005. Distinct regulation of Ubc13 functions by the two ubiquitin-conjugating enzyme variants Mms2 and Uev1A. *J Cell Biol.* 170(5): pp.745-755.

Aridor, M. and Hannah, L.A. 2002. Traffic Jam II: an update of diseases of intracellular transport. *Traffic.* 3: pp.781-790.

Armon, A., Graur, D. and Ben-Tal, N. 2001. Consurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology.* 307: pp.447-463.

Atsumi, A., Tomita, A., Kiyoi, H. and Naoe, T. 2006. Histone deacetylase 3 (HDAC3) is recruited to target promoters by PML-RARα as a component of the N-CoR co-repressor complex to repress transcription in vivo. *Biochem. and Biophy. Res. comm.* 345: pp.1471-1480.

Avvakumov, G.V., Walker, J.R., Xue, S., Finerty, P.J., Mackenzie, F., Newman, E.M. and Dhe-Paganon, S. 2006. [Epub ahead of print]. Amino-terminal dimerization, NRDP1 (FLRF) – rhodanese interaction, and inhibited catalytic domain conformation of the ubiquitin specific protease 8 (USP8/UBPY). *J Biol Chem.*

Baboshina, O.V. and Haas, A.L. 1996. Novel multiubiquitin chain linkages catalyzed by the conjugating enzymes E2EPF and RAD6 are recognized by 26 S proteasome subunit 5. *J Biol Chem.* 271(5): pp.2823-2831.

Baldauf, S.L. 2003. Phylogeny for the faint of heart: a tutorial. *Trends Genet.* 19(6): pp.345-351.

Banerjee, A., Gregori, L., Xu, Y. and Chau, V. 1993. The bacterially expressed yeast CDC34 gene product can undergo autoubiquitination to form a multiubiquitin chain-linked protein. *J Biol Chem.* 268(8): pp.5668-5675.

Bartlett, G.J., Todd, A.E. and Thornton, J.M. 2003. Inferring protein function from structure. In: Bourne, P.E. and Weissig, H. eds. *Structural Bioinformatics*. New Jersey: Wiley. pp.387-407.

Barton, G. 1996. Protein sequence alignment and database scanning. In: Sternberg, M.J.E. ed. *Protein structure prediction a practical approach.* London: Oxford University Press. pp.31-63.

BBSRC Bioscience IT Services. 2003. *Protein Structure and Molecular Modelling.* Herts: BBSRC.

Bence, N.F., Sampat, R.M. and Kopito, R.R. 2001. Impairment of the ubiquitin-proteasome system by protein aggregation. *Science.* 292(5521): pp.1552-1555.

Ben-Tal, N. 2005. *The consurf server.* [online]. Available from: http://consurf.tau.ac.il/ [Accessed 11 March 2006]

Ben-Tal, N. 2005. *The consurf overview.* [online]. Available from: http://consurf.tau.ac.il/overviewver3.html [Accessed 11 March 2006]

Berezin, C., Glaser, F., Rosenberg, Y., Paz, I., Pupko, T., Fariselli, P., Casadio, R., and Ben-Tal, N. 2003. *The ConSeq Server.* [online]. Available from: http://conseq.bioinfo.tau.ac.il/ [Accessed 11 March 2006]

Beynon, R.J. and Bond, J.S. 1989. *Proteolytic enzymes, A Practical Approach.* Oxford: Oxford University Press.

Biederer, T., Volkwein, C. and Sommer, T. 1996. Degradation of subunits of the Sec61p complex, an integral component of the ER membrane, by the ubiquitin-proteasome pathway. *The EMBO Journal.* 15: pp.2069-2076.

Biophysics.org. 2004. *Protein Structure Prediction.* [online]. Available from: http://www.biophysics.org/education/ellis.pdf [Accessed 29 March 2005].

BIO-RAD. [No date] *Bio-Rad protein assay.* [Online]. Available from: http://www.fhcrc.org/science/labs/hahn/methods/biochem_meth/biorad_assay.pdf [Accessed 02 August 2005]

Blow, D. 2002. Outline of crystallography for biologists. New York: Oxford University Press. pp.82-97.

Bonifacino, J.S. and Weissman, A.M. 1998. Ubiquitin and the control of protein fate in the secretory and endocytic pathways. *Annu. Rev. Cell Dev. Biol.* 14: pp.19–57.

Bourne, P.E. and Shindyalov, I.N. 2003, Structure comparison and alignment. In: Bourne, P.E. and Weissig, H. eds. *Structural Bioinformatics.* New Jersey: Wiley. pp. 321-337.

Briffeuil, P., Baudoux, G., Lambert, C., De Bolle, X., Vinals, C., Feytmans, E. and Depiereux, E. 1998. Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions. *Bioinformatics.* 14(4): pp.357-366.

Briggman, K.B., Majumdar, A., Coleman, C.S., Chau, V. and Tolman, J.R. 2005. NMR assignment of human ubiquitin conjugating enzyme Ubc7. *J Biomol NMR.* 32(4):p.340.

Brodsky, J.L. and McCracken, A.A. 1999. ER protein quality control and proteasome-mediated protein degradation. *Cell and Developmental Biology.* 10: pp.507-513.

Brown, T. A. 2002. *Genomes.* 2nd. ed. Manchester: Bios scientific publishers Ltd. pp.108-121.

Brownell, J.E. and Allis C.D. 1995. An activity gel assay detects a single, catalytic active histone acetyltransferase subunit in Tetrahymena macronuclei. *Proc. Natl. Acad. Sci.* 92. pp.6364- 6368.

Bruno, W.J., Socci, N.D. and Halpern, A.L. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol Evol.* 17(1): pp.189-97.

Brzovic, P.S., Lissounov, A., Christensen, D.E., Hoyt, D.W. and Klevit, R.E. 2006. A UbcH5/ubiquitin noncovalent complex is required for processive BRCA1-directed ubiquitination. *Mol Cell.* 21(6): pp.873-880.

Buchwald, G., van der Stoop, P., Weichenrieder, O., Perrakis, A., van Lohuizen, M. and Sixma, T.K. 2006. Structure and E3-ligase activity of the Ring-Ring complex of polycomb proteins Bmi1 and Ring1b. *EMBO J.* 25(11): pp.2465-2474.

Burley, S.K. and Bonanno, J.B. Structural genomics. J.B. 2003. In: Bourne, P.E. and Weissig, H. eds. *Structural Bioinformatics*. New Jersey: Wiley. pp.591-612.

Buschhorn, B.A., Kostova, Z., Medicherla, B., Wolf, D.H. 2004. A genome-wide screen identifies Yos9p as essential for ER- associated degradation of Glycoproteins. *FEBS Letters*. 577(3): pp.422-426.

Calderon, M. and Baumann, W. J. 1970. Gel permeation chromatography of neutral hydroxy lipids on Sephadex LH-20. *Journal of Lipid Research.* 11(March): pp.167-169.

Cancer research UK. 2002. *3D-JIGSAW.* [online]. Available from: http://www.bmm.icnet.uk/servers/3djigsaw/ [Accessed 12 January 2006]

Chau, V., Tobias, J.W., Bachmair, A., Marriott, D., Ecker, D.J., Gonda, D.K. and Varshavsky, A. 1989. A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. *Science.* 243(4898):pp.1576-1583.

Chen, Y., Liu, W., McPhie, D.L., Hassinger, L. and Neve, R.L. 2003. APP-BP1 mediates APP-induced apoptosis and DNA synthesis and is increased in Alzheimer's disease brain. *J Cell Biol.* 163(1): pp.27-33.

Chiba, T. 2005. In vitro systems for NEDD8 conjugation by Ubc12. *Methods Enzymol.* 398: pp.68-73.

Chinea, G., Padron, G., Hooft, R.W.W., Sander, C. and Vriend, G. 1995. The use of position specific rotamers in model building by homology. *Protiens.* 23: pp. 415-421.

Choi, H.J., Zilles, K., Mohlberg, H., Schleicher, A., Fink, G. R. and Armstrong, E. 2006. Cytoarchitectonic identification and probabilistic mapping of two distinct areas within the anterior ventral bank of the human intraparietal sulcus.
J Comp Neurol. 495(1): pp.53-69.

Chor, B. and Tuller, T. 2005. Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics.* 21: pp.i97–i106.

Chothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 105: pp.1-12.

Chotia, C. and Lesk, A.M. 1986. The relation between the divergence and the structure in proteins. *EMBO J.* 5: pp. 823-836.

Ciechanover, A. 1994. The ubiquitin-proteasome proteolytic pathway. *Cell.* 79(October): pp.13-21.

Ciechanover, A. and Schwartz, A.L. 2002. Ubiquitin-mediated degradation of cellular protein in health and disease. *Hepatology.* 35(1): pp.3-6.

Circular Dichroism Spectroscopy of Biomolecules. [online]. Available from: http://www.newark.rutgers.edu/chemistry/grad/chem585/lecture1.html [Accessed 12 September 2005].

Clark T.W., 1997. *Programming issues for Molecular Dynamics.* [online]. Available from: http://amrit.ittc.ku.edu/tclark/bio97/node1.html, http://www.igc.ethz.ch/gromos-docs/index.html [Accessed 18 July 2006]

Clegg, W. 1998. *Crystal Structure Determination.* New York: Oxford University Press.

Connolly, M.L. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science.* 221: pp.709-713.

Cook, W.J., Jeffrey, L.C., Sullivan, M.L. and Vierstra, R.D. 1992. Three-dimensional structure of a ubiquitin-conjugating enzyme (E2). *J Biol Chem.* 267(21): pp.15116-15121.

Cook, W.J., Jeffrey, L.C., Xu, Y. and Chau, V. 1993. Tertiary structures of class I ubiquitin-conjugating enzymes are highly conserved: crystal structure of yeast Ubc4. *Biochemistry.* 32(50): pp.13809-13817.

Cook, W.J., Martin, P.D., Edwards, B.F., Yamazaki, R.K. and Chau, V. 1997. Crystal structure of a class I ubiquitin conjugating enzyme (Ubc7) from Saccharomyces cerevisiae at 2.9 angstroms resolution. *Biochemistry.* 36(7): pp.1621-1627.

Cornell, W.D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. Jr., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. 1995. A second generation force field for the simulation of Proteins, Nucleic acids, and Organic Molecules. *J. Am. Chem. Soc.* 117: pp.5179-5197.

Da Silva, A.E.A., Villanueva, W.J.P., Knidel, H., Bonato, V., Dos Reis, S.F. and Von Zuben, F.J. 2005. A multi-neighbor-joining approach for phylogenetic tree reconstruction and visualization. *Genetics and Molecular Research.* 4(3): pp.525-534.

Davidson College. *Protein Crystallization.* [online]. Available from: http://www.bio.davidson.edu/Courses/Molbio/MolStudents/spring2003/Kogoy/protein.html [Accessed 12 February 2006]

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. 1978. A model of evolutionary change in proteins. matrices for detecting distant relationships. In: Dayhoff, M.O., ed. *Atlas of protein sequence and structure.* Vol 5, Washington DC: National biomedical research foundation. pp.345-358.

Deepview tutorial. [online]. Available from: http://www.usm.maine.edu/~rhodes/SPVTut/index.html [Accessed 21 October 2005]

DeLucas, L.J., Bray, T.L., Nagy, L., McCombs, D., Chernov, N., Hamrick, D., Cosenza, L., Belgovskiy, A., Stoops, B. and Chait, A. 2003. Efficient protein crystallization. *Journal of Structural Biology.* 142: pp.188-206.

Desterro, J.M., Thomson, J. and Hay, R.T. 1997. Ubch9 conjugates SUMO but not ubiquitin. *FEBS Lett.* 417(3): pp.297-300.

Diwan, J. 1998-2006. Protein degradation. Molecular Biochemistry II. [online]. Available from: http://www.rpi.edu/dept/bcbp/molbiochem/MBWeb/mb2/part1/protease.htm [Accessed 11 October 2006]

Dolz, R. 1994. GCG: production of multiple sequence alignment. *Methods Mol Biol.* 24: pp.83-99.

Drake, A. F. 2001. Circular dichroism. In: Harding, S.E. and Chowdhry, B.Z., eds. *Protein-ligand Interactions: structure and spectroscopy Practical approach.* New York: Oxford University press. pp.123-167.

Drenth, J. 1999. *Principles of Protein X-ray Crystallography.* 2nd ed. New York: Springer.

Dunbrack, R.L. Jr. and Karplus, M. 1994. Conformational analysis of the backbone dependent rotamer preferences of protein side chains. *Nat. Struct. Biol.* 5: pp.334-340.

Eckert, J.H. and Johnsson, N. 2003. Pex10p links the ubiquitin conjugating enzyme Pex4p to the protein import machinery of the peroxisome. *J Cell Sci.* 116(Pt 17): pp.3623-3634.

Eddy, S.R. 2004. Where the BLOSUM62 alignment score matrix come from? *Nature Biotechnology.* 22(8): pp.1035-1036.

Eidhammer, I., Jonassen, I. and Taylor, W.R. 2004. *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis.* [s.l.]: John Wiley & Sons, Inc. p.336.

EMBL-EBI. EMBOSS Pairwise Alignment Alignments. [online]. Available from: http://www.ebi.ac.uk/emboss/align/index.html. [Accessed 14 May 2005].

Enabling Grids for E-sciencE. [online]. Available from: http://wisdom.eu-egee.fr/malaria/ [Accessed 10 September 2006]

Ensembl S. cerevisiae. 2006. [online]. Available from: http://www.ensembl.org/Saccharomyces_cerevisiae/index.html [Accessed 21 January 2006]

European Bioinformatics Institute. *Colouring schemes in the M.S.A.* [online]. Available from: http://www.ebi.ac.uk/clustalw/# [Accessed 21 February 2006].

Felsenstein, J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution.* 39(4): pp.783-791

Felsenstein, J. 1986-2004. DRAWTREE and DRAWGRAM. [online]. Available from: http://evolution.genetics.washington.edu/phylip/doc/draw.html [Accessed 5 December 2003]

Felsenstein, J. 1991-2004. NEIGHBOR -- Neighbor-Joining and UPGMA methods. [online]. Available from: http://evolution.genetics.washington.edu/phylip/doc/neighbor.html [Accessed 27 November 2003]

Felsenstein, J. 1991-2004. SEQBOOT -- Bootstrap, Jackknife, or Permutation Resampling of Molecular Sequence, Restriction Site, Gene Frequency or Character Data. [online]. Available from: http://evolution.genetics.washington.edu/phylip/doc/seqboot.html [Accessed 30 October 2003]

Felsenstein, J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J Mol Evol.* 53(4-5): pp.447-55.

Felsenstein, J. 2004. *Inferring Phylogenies.* Massachusetts: Sinauer Associates, Inc. pp.147-175, 335-363, 521-538.

Felsenstein, J. [No date]. *Phylip.* [online]. Available from: http://evolution.genetics.washington.edu/phylip.html [Accessed 5 December 2003].

Fiser, A., Do, R.K. and Šali, A. 2000. Modeling of loops in protein structures. *Protien Sci.* 9: pp.1753-1773.

Flannery, A.V., Beynon, R.J. and Bond, J.S. 1989. Proteolysis of proteins for sequence analysis and peptide mapping. In: Beynon, R.J. and Bond, J.S. eds. *Proteolytic enzyme, a practical approach.* Oxford: IRL Press. pp.145-162.

Fortier, J.M. and Kornbluth, J. 2006. NK lytic-associated molecule, involved in NK cytotoxic function, is an E3 ligase. *J Immunol.* 176(11): pp.6454-6463.

Frasor, J., Danes, J.M., Funk, C.C. and Katzenellenbogen, B.S. 2005. Estrogen down regulation of the corepressor N-CoR: Mechanism and implications for estrogen derepression of N-CoR-regulated genes. *Proc Natl Acad Sci* . 102(37): pp.13153-13157.

Friedlander, R., Jarosch, E., Urban, J., Volkwein, C. and Sommer, T. 2000. A regulatory link between ER-associated protein degradation and the unfolded-protein response. *Nature Cell Biology.* 2(7): pp.379-384.

Garnier, J., Gibrat, JF. and Robson, B. 1996.GOR method for predicting protein secondary structure from amino acid sequence. *Methods in Enzymology.* 266: pp.540-553.

Gelman, M.S. and Kopito, R.R. 2002. Rescuing protein conformation: prospects for pharmacological therapy in cystic fibrosis. *J Clin Invest.* 110(11): pp.1591-1597.

Genedoc Multiple sequence alignment. *Colouring scheme according to physiochemical properties.* [online]. Availablefrom: http://www.psc.edu/biomed/genedoc/tutorial.htm [Accessed 20 November 2005]

Gilon, T., Chomsky, O. and Kulka, R.G. 2000. Degradation signals recognized by the Ubc6p-Ubc7p ubiquitin-conjugating enzyme pair. *Mol Cell Biol.* 20(19): pp.7214-7219.

Glaser, F. Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. 2003. ConSurf: Identification of functional regions in proteins by surface mapping of Phylogenetic Information. *Bioinformatics.* 19(1): pp.163-164.

Godzik, A. 2003. Fold recognition methods. In: Bourne, P.E. and Weissig, H. eds. *Structural Bioinformatics*. New Jersey: Wiley. pp.525-546.

Goldberg, A. L. 2003. Protein degradation and protection against misfolded or damaged proteins. *Nature*. 426(December): pp.895-899.

Goldman, N. and Whelan, S. 2000. Statistical tests of gamma-distribution rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* 17(6): pp.975-978.

Goodsell, D. S. and Olson, A. J. 1990. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins: Str. Func. and Genet.* 8: pp.195-202.

Gu, X., Fu, Y-X. and Li, W-H. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology of Evolution.* 12(4): pp.546-557.

Gu, X. and Zhang, J. 1997. A simple method for estimating the parameter of substitution rate variation among sites. *Mol Biol Evol.* 14(11): pp.1106-1113.

Guex, N., Diemand, A., Peitsch, M.C. and Schwede, T. 2006. Deep View Swiss-Pdb Viewer. [online]. Available from: http://www.expasy.ch/spdbv [Accessed 11 October 2005]

Gulick, A.M., Horswill, A.R., Thoden, J.B., Escalante-Semerena, J.C. and Rayment, I. 2002. Pentaerythritol propoxylate: a new crystallization agent and cryoprotectant induces crystal growth of 2-methylcitrate dehydratase. *Acta Crystallogr D Biol Crystallogr.* 589(2): pp.306-309. [online]. Available from:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pubmed [Accessed 19 June 2005].

Gunsteren, V. 2001. *The GROMOS homepage.* [online]. Available from: http://www.igc.ethz.ch/gromos-docs/index.html [Accessed 17 July 2006]

Guterman, A. and Glickman, M.H. 2004. Complementary roles for Rpn11 and Ubp6 in deubiquitination and proteolysis by the proteasome. *The Journal of Biological Chemistry.* 279(3): pp.1729-1738.

Haas, A.L., Bright, P.M. and Jackson, V.E. 1988. Functional diversity among putative E2 Isozymes in the mechanism of ubiquitin-histone ligation. *The Journal of Biological Chemistry.* 263(26): pp.13268-13275.

Hamilton, K.S., Ellison, M.J., Barber, K.R., Scott Williams, R., Huzil, J.T., McKenna, S., Ptak, C., Glover, M. and Shaw, G.S. 2001. Structure of a conjugating enzyme-ubiquitin thiolester intermediate reveals a novel role for the ubiquitin tail. *Structure.* 9(October): pp.897-904.

Hampton research. *Crystal screen.* [online]. Available from: http://www.hamptonresearch.com/assets/products/attachments/0000000001-0000000073.pdf [Accessed 12August 2005].

Hampton research. *Crystal growth 101 literature.* [online]. Available from: *http://www.hamptonresearch.com/support/Growth101Lit.aspx* [Accessed 12 August 2005].

Hartmann, M. and Golding, G.B. 1998. Searching for substitution rate heterogeneity. *Mol Phylogenet Evol.* 9(1): pp.64-71.

Hay, R.T. 2005. SUMO: a history of modification. *Molecular Cell.* 18(April): pp.1-12.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci.* 89: pp.10915-10919.

Hershko, A. and Ciechanover, A. 1992. The ubiquitin system for protein degradation. *Annu. Rev. Biochem.* 61: pp.761-807.

HGNC. [No date] *UBE2 search results*. [online]. Available from: http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl [Accessed 21 January 2006]

Hisatake, K., Hasegawa, S., Takada, R., Nakatani, Y., Horikoshi, M. and Roeder, R.G. 1993. The p250 subunit of native TATA box-binding factor TFIID is the cell-cycle regulatory protein CCG1. *Nature.* 362(March): pp. 179-181.

Hoppe, T. 2005. Multiubiquitylation by E4 enzymes: 'one size' doesn't fit all. Trends Biochem Sci. 30(4): pp.183-187.

Horie-Inoue, K. and Inoue, S. 2006. Epigenetic and proteolytic inactivation of 14-3-3sigma in breast and prostate cancers. *Semin Cancer Biol.* 16(3): pp.235-239.

Houben K, Dominguez C, van Schaik FM, Timmers HT, Bonvin AM, Boelens R. 2004. Solution structure of the ubiquitin-conjugating enzyme UbcH5B. *J. Mol. Biol.* 344: pp.513-526.

Howe, K., Bateman, A. and Durbin, R. 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics.* 18(11): pp.1546-1547.

Huang, D.T., Paydar, A., Zhuang, M., Waddell, M.B., Holton, J.M. and Schulman, B.A. 2005. Structural basis for recruitment of Ubc12 by an E2 binding domain in NEDD8's E1. *Mol Cell.* 17(3): pp.341-350.

Huang, D.T., Miller, D.W., Mathew, R., Cassell, R., Holton, J.M., Roussel, M.F. and Schulman, B.A. 2004. A unique E1-E2 interaction required for optimal conjugation of the ubiquitin-like protein NEDD8. *Nature Structural & Molecular Biology.* 11(10): pp.927-935.

Huang, J., Xu, L.G., Liu, T., Zhai, Z. and Shu, H.B. 2006. The p53-inducible E3 ubiquitin ligase p53RFP induces p53-dependent apoptosis. *FEBS Lett.* 580(3): pp.940-947.

Huang, L., Kinnucan, E., Wang, G., Beaudenon, S., Howley, P.M., Huibregtse, J.M. and Pavletich, N.P. 1999. Structure of an E6AP-ubcH7 complex: Insights into ubiquitination by the E2-E3 enzyme cascade. *Science.* 286(November): pp.1321-1325.

HUGO Gene Nomenclature Committee. [online]. Available from: http://www.gene.ucl.ac.uk/nomenclature/ [Accessed 09 July2005]

Hung, L-H. and Samudrala, R. 2003. PROTINFO: secondary and tertiary protein structure prediction. *Nucleic Acids Research.* 31(13): pp.3296-3299.

Huyer, G., Piluek, W.F., Fansler, Z., Kreft, S.G., Hochstrasser, M., Brodsky, J.L. and Michaelis, S. 2004. Distinct machinery is required in saccharomyces cerevisiae for the Endoplasmic Reticulum-associated Degradation of a Multispanning Membrane protein and a soluble luminal Protein. *J. Biol. Chem.* 279(37): pp.38369-38378.

Indiana University. 2001. *Insight II notes: Forcefields.* [online]. Available from: http://www.molvis.indiana.edu/app_guide/InsightII/forcefields.html [Accessed 17 July 2006]

Irwin JJ and Shoichet BK. 2005. ZINC - A free database of commercially available compounds for virtual screening *J. Chem. Inf. Comput.* Sci. 45(1): pp.177-182.

Jacobson, R.H., Ladurner, A.G., King, D.S. and Tjian, R. 2000. Structure and function of a human TAF_{II} 250 double bromodomain module. *Science.* 288(May): pp.1422-1425.

Jentsch, S. 1992. The Ubiquitin-Conjugation System. *Annual Review of Genetics.* 26(December): pp.179-207.

Jentsch, S. 2005. [online]. Available from: http://www.biochem.mpg.de/jentsch/Jentsch.html [Accessed 20 October 2005]

Jones, D., Crowe, E., Stevens, T.A. and Candido, E.P. 2001. Functional an Phylogenetic analysis of the ubiquitilation system in Ceanorhabditis elegans:ubiquitin-conjugating enzymes, ubiquitin-activating enzymes, and ubiquitin-like proteins. *Genome Biology.* 3(1): pp.0002.1-0002.15.

Jude Stevens, B.S. and Gary Kobbs, B.S. 2004. FastBreak™ Cell lysis reagent for protein purification. [online]. Available from: http://www.promega.com/pnotes/86/11217_23/11217_23.pdf [Accessed 21 March 2005]

Jungmann, J., Reins, H.A., Schobert, C. and Jentsch, S. 1993. Resistance to cadmium mediated by ubiquitin-dependent proteolysis. *Nature.* 361(6410): pp.369-371.

Kaiser, P. and Huang, L. 2005. Global approaches to understanding ubiquitination. *Genome Biology.* 6(10):p233.

Karaczyn, A.A., Golebiowski, F. and Kasprzak, K.S. 2006. Ni(II) affects ubiquitination of core histones H2B and H2A. *Exp Cell Res.(article in press).*

Khan, M.M., Nomura, T., Chiba, T., Tanaka, K., Yoshida, H., Mori, K. and Ishii, S. 2004. The Fusion Oncoprotein PML-RARα Induces Endoplasmic Reticulum (ER)-associated Degradation of N-CoR and ER Stress. *The Journal of Biological Chemistry.* 279(12): pp.11814-11824.

Kelly, S.M., Jess, T.J. and Price, N.C. 2005. How to study proteins by circular dichroism. *Biochim Biophys Acta.* 1751(2): pp.119-139.

Kloor, M., Bork, P., Duwe, A., Klaes, R., von Knebel Doeberitz, M. and Ridder, R. 2002. Identification and characterization of UEV3, a human cDNA with similarities to inactive E2 ubiquitin-conjugating enzymes. *Biochimica et Biophysica Acta.* 1579: pp.219-224.

Kobirumaki, F., Miyauchi, Y., Fukami, K. and Tanaka, H. 2005. A novel UbcH10-binding protein facilitates the ubiquitinylation of cyclin B in vitro. *J Biochem.* 137(2): pp.133-139.

Koonin, E.V. and Abagyan, R.A. 1997. TSG101 may be the prototype of a class of dominant negative ubiquitin regulators. *Nature Genetics.* 16(August): pp.330-331.

Koshiyama, A., Hamada, F.N., Namekawa, S.H., Iwabata, K., Sugawara, H., Sakamoto, A., Ishizaki, T. and Sakaguchi, K. 2006. Sumoylation of a meiosis-specific RecA homolog, Lim15/Dmc1, via interaction with the small ubiquitin-related modifier (SUMO)-conjugating enzyme Ubc9. *FEBS J.* 273(17): pp.4003-4012.

Kostova, Z. and Wolf, D.H. 2003. For whom the bell tolls: protein quality control of the endoplasmic reticulum and the ubiquitin-proteasome connection. *The EMBO Journal.* 22(10): pp.2309-2317.

Knipscheer, P. and Sixma, T.K. 2006. Divide and conquer: the E2 active site. *Nature Structural & Molecular Biology.* 13(6): pp. 474-476.

Krause, R., Nielsen, J.E. and Vriend, G. 2006. *Whatif web interface.* [online]. Available from: http://swift.cmbi.ru.nl/WIWWWI/ [Accessed 20 January 2006]

Krieger, E., Nabuurs, S.B. and Vriend, G. 2003. Homology modelling. In: Bourne, P.E. and Weissig, H. eds. *Structural Bioinformatics.* New Jersey: Wiley. pp.509-521.

Kuhner, M.K. and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution.* 11: pp.459-468

Kyte, J. and Doolittle, R. 1982. Kyte-Doolittle Hydropathy Plots. [online]. Available from: http://occawlonline.pearsoned.com/bookbind/pubbooks/bc_mcampbell_genomics_1/medialib/activities/kd/kyte-doolittle-background.htm#plot [Accessed 3 September 2005].

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. and Ben-Tal, N. 2005. Consurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids research.* 33: pp.W299-W302.

Lederkremer, G.Z. and Glickman, M.H. 2005. A window of opportunity: timing protein degradation by trimming of sugars and ubiquitins. *Trends in Biochemical Sciences.* 30(6): pp.297-303.

Lee, B.K. and Richards, F.M. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55: pp.379-400.

Lee, S.J., Choi, J.Y., Sung, Y.M., Park, H., Rhim, H. and Kang, S. 2001. E3 ligase activity of RING finger proteins that interact with Hip-2, a human ubiquitin-conjugating enzyme. *FEBS Lett.* 503(1): pp.61-64.

Rarey, M., Kramer, B., Lengauer, T. and Klebe, G. 1996. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol.* 261(3): pp.470-489.

Lenk, U., Yu, H., Walter, J., Gelman, M.S., Hartmann, E., Kopito, R.R. and Sommer, T. 2002. A role for mammalian Ubc6 homologues in ER-associated protein degradation. *Journal of Cell Science.* 115: pp.3007-3014.

Lester, D., Farquharson, C., Russell, G. and Houston, B. 2000. Identification of a family of noncanonical ubiquitin-conjugating enzymes structurally related to yeast UBC6. *Biochem Biophys Res Commun.* 269(2): pp.474-480.

Lewis, M.J., Saltibus, L.F., Hau, D.D., Xiao, W. and Spyracopoulos, L. 2006. Structural basis for non-covalent interaction between ubiquitin and the ubiquitin conjugating enzyme variant human MMS2. *J Biomol NMR.* 34(2): pp.89-100.

Li, L. and Cohen, S.N. 1996. Tsg101: a novel tumor susceptibility gene isolated by controlled homozygous functional knockout of allelic loci in mammalian cells. *Cell.* 85(3): pp.319-329.

Li, L., Li, X., Francke, U. and Cohen, S.N. 1997. The TSG101 tumor susceptibility gene is located in chromosome 11 band p15 and is mutated in human breast cancer. *Cell.* 88(1): pp.143-154.

Lin, H. and Wing, S.S. 1999. Identification of rabbit reticulocyte $E2_{17k}$ as a UBC homologue and functional characterization of its core domain loop. *The journal of Biological Chemistry.* 274(21): pp.14685-14691

Linding, R. 2003. Globplot 2. Intrinsic Protein Disorder, Domain & Globularity Prediction. [online]. Available from: http://globplot.embl.de/ [Accessed 19 September 2005]

Linding, R., Russell, R.B., Neduva, V. and Gibson, T.J. 2003. Globplot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research.* 31(13): pp.3701-3708.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J. and Russell, R.B. 2003. Protein disorder prediction: implications for structural proteomics. *Structure.* 11(11): pp.1453-1459.

Linding, R. 2004. *Help on using Globplot.* [online]. Available from: http://globplot.embl.de/html/help.html [Accessed 19 September 2005]

Lio P and Goldman N. 1998. Models of molecular evolution and Phylogeny. *Genome Research.* 8: pp.1233- 1244

Liu, Q., Shang, F., Whitcomb, E., Guo, W., Li, W. and Taylor, A. 2006. Ubiquitin-conjugating enzyme 3 delays human lens epithelial cells in metaphase. *Invest Ophthalmol Vis Sci.* 47(4): pp.1302-1309.

Liu, Z., Oughtred, R. and Wing, S.S. 2005. Characterization of E3Histone, a novel testis ubiquitin protein ligase which ubiquitinates histones. *Molecular and Cellular Biology.* 25(7): pp.2819-2831.

Lobley, A., Whitmore, L. and Wallace, B.A. 2002. DICHROWEB: an interactive website for the analysis of protein secondarystructure from circular dichroism spectra. *Bioinformatics*. 18(1): pp.211-212.

Lopez-Avalos, M.D., Duvivier-Kali, V.F., Xu, G., Bonner-Weir, S., Sharma, A. and Weir, G.C. 2006. Evidence for a role of the ubiquitin-proteasome pathway in pancreatic islets. *Diabetes*. 55(5): pp.1223-1231.

MacKerell A.D., [No date] *Protein Force Fields*. [online]. Available from: http://www.wiley.com/legacy/wileychi/ecc/samples/sample07.pdf [Accessed 17 July 2006]

Mancini, R., Aebi, M. and Helenius, A. 2003. Multiple Endoplasmic Reticulum-associated pathways degrade mutant yeast carboxypeptidase Y in mammalian cells. *J. Biol. Chem.* 278(47): pp.46895-46905.

Martin, P.D., Edwards, B.F.P., Yamazaki, R.K. and Chau, V. 1997. Crystal structure of a class I ubiquitin conjugating enzyme (UBC7) from saccharomyces cerevisiae at 2.9Å resolution. *Bochemistry*. 36: pp.1621-1627.

McCracken, A.A. and Brodsky L. J. 2003. Evolving questions and paradigm shifts in endoplasmic-reticulum associated degradation (ERAD). *Bioessays*. 25: pp.868-877.

McPherson, A. 2003. Introduction to protein crystallization. *Methods*. 34(3): pp.254-265.

McPherson, A. 1999. *Crystallization of Biological Macromolecules*. New York: Cold Spring Harbor Laboratory Press.

McRee, Duncan E. 1993. *Practical Protein Crystallography*. San Diego: Academic Press. pp.1-23.

Meusser, B., Hirsch, C., Jarosch, E. and Sommer, T. 2005. ERAD: the long road to destruction. *Nature Cell Biology*. 7: pp.766-772.

Michalsky, E., Goede, A. and Preissner, R. 2003. Loops in proteins (LIP)- a comprehensive loop database for homology modelling. *Protein Engineering.* 16(12): pp.979-985.

Mira, M.T., Alcais, A., Nguyen, V.T., Moraes, M.O., Di Flumeri, C., Vu, H.T., Mai, C.P., Nguyen, T.H., Nguyen, N.B., Pham, X.K., Sarno, E.N., Alter, A., Montpetit, A., Moraes, M.E., Moraes, J.R., Dore, C., Gallant, C.J., Lepage, P., Verner, A., Van De Vosse, E., Hudson, T.J., Abel, L. and Schurr, E. 2004. Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature.* 427(6975): pp.636-640.

Mizzen, C.A. and Allis, C.D. 2000. New insights into an old modification. *Science.* 289(September): pp.2290-2291.

Mo, Y-Y and Moschos, S.J. 2005. Targetting Ubc9 for cancer therapy. *Exper Opinion on Therapeutic Targets.* 9(6): pp.1203-1216.

Morris, G.M., Goodsell, D.S., Huey, R. and Olson, A.J. 1996. Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Computer-Aided Molecular Design.* 10: pp.293-304.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K. and Olson, A.J. 1998. Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function. *J. Computational Chemistry.* 19: pp.1639-1662.

Morris, G.M. [No date]. *Auto Dock: Automated docking of flexible ligands.* [online]. Available from: http://www.scripps.edu/mb/olson/doc/autodock/ [Accessed 05 September 2006]

Nagl, S. 2003. Molecular evolution. In: Orengo, C., Jones, D. and Thronton, J. eds. *Bioinformatics- Genes, Proteins & Computers.* Oxford: BIOS Scientific Publishers Limited, pp.65-80.

Nameki, N., Yoneyama, M., Koshiba, S., Tochio, N., Inoue, M., Seki, E., Matsuda, T., Tomo, Y., Harada, T., Saito, K., Kobayashi, N., Yabuki, T., Aoki, M., Nunokawa, E., Matsuda, N., Sakagami, N., Terada, T., Shirouzu, M., Yoshida, M., Hirota, H., Osanai, T., Tanaka, A., Arakawa, T., Carninci, P., Kawai, J., Hayashizaki, Y., Kinoshita, K., Guntert, P., Kigawa, T. and Yokoyama, S. 2004. Solution structure of the RWD domain of the mouse GCN2 protein. *Protein Science.* 13(8): pp.2089-2100.

National Grid Service. 2006. [online]. Available from: http://www.grid-support.ac.uk/ [Accessed 05 September 2006].

NCBI Glossary. [online]. Available from: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html [Accessed 12 August 2005].

Needleman, S. B. and Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48: pp.443-453

Notredame, C., Higgins, D. and Heringa, J. 2000. T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology*, 302: pp.205-217.

Nuber, U., Schwarz, S., Kaiser, P., Schneider, R. and Scheffner, M.. 1996. Cloning of human ubiquitin conjugating enzymes ubcH6 and ubcH7 (E2-F1) and characterization of their interaction with E6-AP and RSP5. *J Biol Chem.* 271(5): pp.2795-2800.

Ogawa, S., Lozach, J., Jepsen, K., Sawka-Verhelle, D., Perissi, V., Sasik, R., Rose, D.W., Johnson, R.S., Rosenfeld, M.G. and Glass, C.K. 2004. A nuclear receptor corepressor transcriptional checkpoint controlling activator protein 1-dependent gene networks required for macrophage activation. *PNAS.* 101(40): pp.14461-14466.

Oh R.S., Bai, X. and Rommens, J.M. 2006. Human homologs of Ubc6p ubiquitin-conjugating enzyme and phosphorylation of HsUbc6e in response to endoplasmic reticulum stress. *J Biol Chem.* 281(30): pp.21480-21490.

Ohio Supercomputer Center. *E Value in BLAST.* [online]. Available from: http://www.osc.edu/research/bioinformatics/FAQ/evalue.shtml [Accessed 21 May 2005].

OPAL. [online]. Available from: http://www.oppf.ox.ac.uk/opal/OPAL.php?POSTNUKESID=3e72bb35034661e1b0c67007905036ed [Accessed 19 September 2005]

Page R.D.M. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Cabios Applications Note.* 12(4): pp.357-358.

Palmer, R.A. 2001. X-ray crystallographic studies of protein-ligand interactions . In: Harding, S.E. and Chowdhry, B.Z., eds. *Protein-ligand Interactions: structure and spectroscopy Practical approach.* New York: Oxford University press. pp. 29-31.

Passmore, L.A. and Barford, D. 2004. Getting into position: the catalytic mechanisms of protein ubiquitilation. *Biochem J.* 379: pp.513-525.

Patthy, L. 1999. Evolution of protein coding genes. In: *Protein evolution.* Hungary: Blackwell science Ltd. pp.51-74.

Pearson, W.R. 1997. Identifying distantly related potein sequences. *Cabios Invited Review.* 13(4): pp.325-332.

Peitsch, M.C. 2002. About the use of protein models. *Bioinformatics.* 18(7): pp.934-938.

Pham, A.D. and Sauer, F. 2000. Ubiquitin-activating/conjugating activity of TAFII250, a mediator of activation of gene expression in Drosophila. *Science.* 289(5488): pp.2357-2360.

Pickart, C.M. 2001. Mechanisms underlying ubiquitination. *Annu Rev Biochem.* 70: pp.503-533.

Pierce Biotechnology, Inc. 2002. *Cell lysis using detergents.* [online]. Available from: http://www.piercenet.com/Objects/view.cfm?type=Page&ID=1904ED25-8FA4-475C-8068-C2EB13D5F4E7 [Accessed 12 April 2005].

Plemper, R.K. and Wolf, D.H. 1999. Retrograde protein translocation: ERADication of secretory proteins in health and disease. *Trends Biochem Sci.* 24(7): pp.266-270.

Poirot, O., O'Toole, E. and Notredame, C. 2003. Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res.* 31(13): pp.3503-3506.

Pole Bioinformatique Lyonnais.[online]. Available from: http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html [Accessed 2 August 2006].

Ponting, C.P., Cai, Y.D. and Bork, P. 1997. The breast cancer gene product TSG101: a regulator of ubiquitination? *J Mol Med.* 75(7): pp.467- 469.

Project 3-Basic Cloning and Expression- MalE -AAP'1 (YHR047c) fusion in bacterial expression vector. [online]. Available from: http://csm.colostate-pueblo.edu/biology/dcaprio/351L/Project3S06.htm) [Accessed 2 February 2005]

Promega. MagneGST™ Protein Purification System. [online]. Available from: http://www.promega.com/tbs/tm240/tm240.pdf [Accessed 23 June 2005]

Ptak, C., Prendergast, J.A., Hodgins, R., Kay, C.M., Chau, V. and Ellison, M.J. 1994. Functional and physical characterisation of the cell cycle ubiquitin-conjugating enzyme CDC34 (UBC3). *The Journal of Biological Chemistry.* 269(42): pp.26539-26545.

Qiagen. 2002. *QIAquick spin handbook.* [online]. Available from: (http://www1.qiagen.com/literature/handbooks/PDF/DNACleanupAndConcentration/QQ_Spin/1021422_HBQQSpin_072002WW.pdf [Accessed 23 April 2005]

QIAGEN. 2001. Ni-NTA Magnetic Agarose Beads handbook. [online]. Available from: http://www1.qiagen.com/literature/handbooks/PDF/Protein/Purification/NiNTA_Magnetic_Agarose_Beads/1018847HBNiNTAMB_1201WW.pdf [Accessed 10 June 2005]

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. 1963. Ramachandran plot. Stereochemistry of Polypeptide Chain Conformations. *J. Mol. Biol.* 7: pp.95-99.
Rhodes, G. 1993. *Crystallography Made Crystal Clear.* San Diego: Academic Press. pp. 8-10, 29-38.

Rhodes, G. 2000. *Crystallography Made Crystal Clear.* San Diego: Academic Press. pp. 13-14, 46, 50-52.

Rhodes, G. 2000. *Molecular modelling for beginners. Tutorial for DeepView.* [online]. Available from: http://www.usm.maine.edu/~rhodes/SPVTut/index.html [Accessed 2 March 2006].

Rice, P., Longden, I. and Bleasby,A. 2000. *EMBOSS: The European Molecular Biology Open Software Suite.* Trends in Genetics 16, (6): pp.276-277.

Richly, H., Rape, M., Braun, S., Rumpf, S., Hoege, C. and Jentsch, S. 2005. A series of ubiquitin binding factors connects CDC48/p97 to substrate multiubiquitilation and proteasomal targeting. *Cell.* 120(January): pp.73-84.

Riordan JR. 1999. Cystic fibrosis as a disease of misprocessing of the cystic fibrosis transmembrane conductance regulator glycoprotein. *Am J Hum Genet.* 64(6): pp.1499-1504.

Rodriguez, R., Chinea, G., Lopez, N., Pons, T. and Vriend, G. 1998. Homology modelling, model and software evaluation: three related resources. *Bioinformatics.* 14(6): pp.523-528.

RONN. 2005. [online] Available from: http://www.strubi.ox.ac.uk/cgi-bin/disorder_results_jan2005.cgi [Accessed 19 September 2005]

Rose, J.P. and Wang, B-C. 1997. X-stream™ cryocrystallography. *The Rigaku Journal.* 14(1): pp.4-12.

Rosenberg, M.S. 2005. Multiple sequence alignment accuracy and evolutionary distance estimation. *BMC Bioinformatics.* 6(November): p.278.

Rosenberg, M.S. 2005. Evolutionary distance estimation and fidelity of pair wise sequence alignment. *BMC Bioinformatics.* 6(102): pp.1-9.

Roth, A.F. and Davis N.G. 1996. Ubiquitination of the yeast a-factor receptor. *J Cell Biol.* 134(3): pp.661-674.

Rothofsky, M. L. and Lin, S. L. 1997. CROC-1 encodes a protein which mediates transcriptional activation of the human FOS promoter. *Gene.* 195(2): pp.141-149.

Ruppert, S. Wang, E.H. and Tjian, R. 1993. Cloning and expression of human $TAF_{II}250$: a TBP-associated factor implicated in cell-cycle regulation. *Nature.* 362(March): pp. 175-178.

Rutishauser, F. and Spiess, M. 2002. Endoplasmic reticulum storage diseases. *Swiss Med Wkly*; 132: pp.211-222.

Saitou and Nei, 1987. Neighbour joining method. [online]. Available from: http://www.tau.ac.il/~doronadi/neighbourJoining.pdf#search=%22Neighbour-joining%20%22. [Accessed 20 April 2004].

Sancho, E., Vila, M.R., Sanchez-Pulido, L., Lozano, J.J., Paciucci, R., Nadal, M., Fox, M., Harvey, C., Bercovich, B., Loukili, N., Ciechanover, A., Lin, S.L., Sanz, F., Estivill, X., Valencia, A. and Thomson, T.M. 1998. Role of UEV-1, an inactive variant of the E2 ubiquitin-conjugating enzymes, in invitro differentiation and cell cycle behaviour of HT-29-M6 intestinal mucosecretory cells. *Molecular and cellular biology.* 18(1): pp.576-589

Sarcevic, B., Mawson, A., Baker, R.T. and Sutherland, R.L. 2002. Regulation of the ubiquitin-conjugating enzyme hHR6A by CDK-mediated phosphorylation. *EMBO J.* 21(8): pp.2009-2018.

Scheeff, E.D. and Fink, J.L. 2003. Fundamentals of protein structure. In: Bourne, P.E. and Weissig, H. eds. *Structural Bioinformatics*. New Jersey: Wiley. pp.15-39.

Schmidt, H.A., Strimmer, K. and Haeseler, A.V. 2004. Tree-puzzle-maximum likelihood analysis for nucleotides, amino acids, and two-state data. [online]. Available from: http://www.dkfz.de/tbi_old/tree-puzzle/tree-puzzle.pdf [Accessed 3 March 2003].

Schwartz, R.M. and Dayhoff, M.O. 1978. In: Dayhoff, M.O., ed. *Atlas of protein sequence and structure*. Vol 5, Washington DC: National biomedical research foundation. pp.353-362.

Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. 2003. Swiss-model: an automated protein homology-modeling server. *Nucleic Acids Research.* 31(13): pp.3381-3385.

Schwede, T., Peitsch, M.C. and Guex, N. [No date]. *Biozentrum.* [online]. Available from: http://www.expasy.org/swissmod/ [Accessed 5 November 2005].

Seufert, W. and Jentsch, S. 1990. Ubiquitin-conjugating enzymes UBC4 and UBC5 mediate selective degradation of short-lived and abnormal proteins. *The EMBO Journal.* 9(2): pp.543-550.

Sigma Aldrich. *Imidazole purification and detection.* [online]. Available from: http://www.sigmaaldrich.com/Area_of_Interest/Life_Science/Proteomics_and_Protein_Expr_/Protein_Expression/Purification_and_Detection/HIS_Select/Imidazole.html [Accessed 05 April 2005]

Sigma Aldrich. *HIS-Select™ Elution Buffer.* [online]. Available from: http://www.sigmaaldrich.com/catalog/search/ProductDetail/SIGMA/H5413 [Accessed 04 April 2005]

Sjolander, K. 2004. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics.* 20(2): pp.170-179.

Soboleva, T.A. and Baker, R.T. 2004. Deubiquitinating enzymes: their functions and substrate specificity. *Curr Protein Pept Sci.* 5(3): pp.191-200.

South African Structural Biology Initiative, *E. coli* Growth and Induction with IPTG. [online]. Available from: http://sbio.uct.ac.za/Sbio/documentation/Cells%20growth%20and%20induction.html [Accessed 25 March 2005]

Sreerama, N. and Woody, R.W. 2000. Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal Biochem.* 287(2): pp.252-260.

Srinavasan, N., Guruprasad, K. and Blundell, T.L. 1996. Comparative modelling of proteins. In: Sternberg, M.J.E. ed. *Protein structure prediction A practical approach.* Oxford: IRL Press. pp.111-140.

Strachan, T. and Read, A.P. 1999. *Human Molecular Genetics.* 2nd ed. Oxford: Bios scientific publishers Ltd.

Steipe, B. 1999.The construction of the Dayhoff matrix. [online]. Available from: http://www.lmb.uni-muenchen.de/groups/bioinformatics/04/ch_04_3.html [Accessed 2 September 2004]

Sternberg, M.J.E. 1996. Protein structure prediction-principles and approaches. In: *Protein structure prediction A practical approach.* Oxford: IRL Press. p.3.

Stum, Z., Kmetko, J., O'Neill, K., Gillilan, R., Bartnick, A. and Thorne, R.E. 2004. A new crystal mounting method for macromolecular cryocrystallography. *Synchrotron Radiation News.* 17(2): pp.31-32.

Takeuchi, T., Iwahara, S., Saeki, Y., Sasajima, H. and Yokosawa, H. 2005. Link between the Ubiquitin Conjugation System and the ISG15 Conjugation System: ISG15 Conjugation to the UbcH6 Ubiquitin E2 Enzyme. *J Biochem.* 138(6): pp.711-719.

Takeuchi, T., Inoue, S. and Yokosawa, H. 2006. Identification and Herc5-mediated ISGylation of novel target proteins. *Biochem Biophys Res Commun.* 348(2): pp.473-477.

Taxis, C., Hitt, R., Park, S.H., Deak, P.M., Kostova, Z. and Wolf, D.H. 2003. Use of modular substrates demonstrates mechanistic diversity and reveals differences in Chaperone requirement of ERAD. *The Journal of Biological Chemistry.* 278(38): pp.35903-35913.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*earch 22(22): pp.4673-4680.

Thompson, J.D., Plewniak, F. and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 27(13): pp.2682-2690.

TIGR. 2005-2006. [online]. Available from: www.tigr.org. [Accessed 23 March 2004].

Tiwari, S. and Weissman, A.M. 2001. Endoplasmic reticulum (ER)-associated degradation of T cell receptor subunits.Involvement of ER-associated ubiquitin-conjugating enzymes (E2s). *J Biol Chem.* 276(19): pp.16193-16200.

Tong, H., Hateboer, G., Perrakis, A., Bernards, R. and Sixma, T.K. 1997. Crystal structure of murine/human Ubc9 provides insight into the variability of the ubiquitin-conjugating system. *J Biol Chem.* 272(34): pp.21381-21387.

Tree construction. [Online] Available from http://trishul.sci.gu.edu.au/courses/ss13bmm/phylogeny_partA.doc [Accessed 9 August 2006]

Tree of life web project. [online]. Available from: http://tolweb.org/tree/home.pages/glossary.html#polyphyletic [Accessed 9 August 2006]

Tripathi, M.K. and Chaudhuri, G. 2005. Down-regulation of UCRP and UBE2L6 in BRCA2 knocked-down human breast cells. *Biochem Biophys Res Commun.* 328(1): pp.43-48.

Tusnady, G.E., Dosztanyi, Z. and Simon, I. 2005. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.* 33: pp.275-278. [online]. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=pubmed [Accessed 20 July 2005].

Uniprot. [online]. Available from: http://www.ebi.uniprot.org/index.shtml [Accessed 17 February 2004].

VanDemark, A.P., Hofmann, R.M., Tsui, C., Pickart, C.M. and Wolberger, C. 2001. Molecular insights into polyubiquitin chain assembly: crystal structure of the Mms2/Ubc13 heterodimer. *Cell.* 105(6): pp.711-720.

Vriend, G. 1990. WHAT IF: A molecular modelling and drug design program. *J. Mol. Graphics.* 8: pp.52-56.

Vriend, G. 1997. [online]. Available from: http://swift.cmbi.ru.nl/gv/pdbreport/checkhelp/ [Accessed 4 March 2006].

Wain, H.M., Lush, M.J., Ducluzeau, F., Khodiyar, V.K. and Povey, S. 2004. Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Research.* 32: pp. 255-257. [online]. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=14681406&query_hl=11&itool=pubmed_docsum [Accessed 6 March 2004]

Wallace, B.A. and Janes, R.W. 2001. Synchrotron radiation circular dichroism spectroscopy of proteins: secondary structure, fold recognition and structural genomics. *Curr Opin Chem Biol.* 5(5): pp.567-571.

Wanker, E.E., Rovira, C., Scherzinger, E., Hasenbank, R., Walter, S., Tait, D., Colicelli, J. and Lehrach, H. 1997. HIP-I: a huntingtin interacting protein isolated by the yeast two-hybrid system. *Hum. Mol. Genet.* 6 (3): pp.487-495.

Ward, C.L., Omura, S. and Kopito, R.R. 1995. Degradation of CFTR by the ubiquitin-proteasome pathway. *Cell.* 83(1): pp.121-127.

Wassarman, D.A., Aoyagi, N., Pile, L.A. and Schlag, E.M. 2000. TAF250 is required for multiple developmental events in Drosophila. *Proc Natl Acad Science.* 97(3): pp.1154-1159.

Wassarman, D.A. and Sauer, F. 2001. TAF(II)250: a transcription toolbox. *Journal of Cell Science.* 114(16): pp.2895-2902.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weinerl, P. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 106: pp.765-784.

Weiner, S. J. 1986 An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comp. Chem.* 7: pp.230-252.

Whelan, S., Li, P. and Goldman, N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *TRENDS in Genetics.* 17(5): pp.262-272.

Whitmore, L. and Wallace, B.A. 2004. DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Research.* 32 (web server issue DOI:10.1093/nar/gkh371): pp.W668-W673.

Wikipedia.[No date]. *Size exclusion chromatography.* [online]. Answers .com. Available from: http://www.answers.com/topic/size-exclusion-chromatography [Accessed 14 September 2006].

Wikepedia, the free encyclopedia. Force Field (Chemistry). [online]. Available from: http://en.wikipedia.org/wiki/Force_field_%28chemistry%29 [Accessed 17 July 2006]

Winn, P.J., Religa, T.L., Battey, J.N.D., Banerjee, A. and Wade, R.C. 2004. Determinants of functionality in the ubiquitin conjugating enzyme family. *Structure.* 12(9): pp.1563-1574.

Wong, J.J.Y., Pung, Y,F., Siu-Kwan Sze, N. and Chin, K-C. 2006. HERC5 is an IFN-induced HECT-type E3 protein ligase that mediates type I IFN-induced ISGylation of protein targets. *Proc Natl Acad Sci U S A.* 103(28): pp.10735-10740.

Woods, Y.L., Xirodimas, D.P., Prescott, A.R., Sparks, A., Lane, D.P. and Saville, M.K. 2004. p14 Arf promotes small ubiquitin-like modifier conjugation of Werners helicase. *J Biol Chem.* 279(48): pp. 50157-50166.

Worthylake, D.K., Prakash, S., Prakash, L. and Hill, C.P. 1998. Crystal structure of the Saccharomyces cerevisiae ubiquitin-conjugating enzyme Rad6 at 2.6 A resolution. *J Biol Chem.* 273(11): pp.6271-6276.

Wu, C.J., Conze, D.B., Li, X., Ying, S.X., Hanover, J.A. and Ashwell, J.D. 2005. TNF-α induced c-IAP1/TRAF2 complex translocation to a Ubc6-containing compartment and TRAF2 ubiquitination. *The EMBO Journal.* 24: pp.1886–1898.

Xiao, W., Lin, S.L., Broomfield, S., Chow, B.L. and Wei, Y.F. 1998. The products of the yeast MMS2 and two human homologs (hMMS2 and CROC-1) define a structurally and functionally conserved Ubc-like protein family. *Nucleic Acids Research.* 26(17): pp.3908-3914.

Yamamoto, M., Okamoto, T., Takeda, K., Sato, S., Sanjo, H., Uematsu, S., Saitoh, T., Yamamoto, N., Sakurai, H., Ishii, K.J., Yamaoka, S., Kawai, T., Matsuura, Y., Takeuchi, O. and Akira, S. 2006. Key function for the Ubc13 E2 ubiquitin-conjugating enzyme in immune receptor signaling. *Nature Immunol.* 7(9): pp.962- 970.

Yang, Z. and Kumar, S. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol Biol Evol.* 13(5): pp.650-659.

Yang, Z.R., Thomson, R., McNeil, P. and Esnouf, R.M. 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics.* 21 (16): pp.3369-3376.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10(6): pp.1396-1401.
Yip, K. 1997. Types of Chromatography. [online]. Available from:
http://www.rpi.edu/dept/chem-eng/Biotech-Environ/CHROMO/be_types.htm
[Accessed 27 July 2005]

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution.* 39: pp.306-314.

Yang, Z. 1996. Maximum-likelihood models for combined analysis of multiple sequence data. *Journal of Molecular Evolution.* 42: pp.587-596.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Tree.* 11(9): pp.367-371.

Yoshida, Y., Chiba, T., Tokunaga, F., Kawasaki, H., Iwai, K., Suzuki, T., Ito, Y., Matsuoka, K., Yoshida, M., Tanaka, K. and Tai, T. 2002. E3 ubiquitin ligase that recognizes sugar chains. *Nature.* 418(6896): pp. 438-442.

Younger, J.M., Ren, H-Y., Chen, L., Fan, C-Y, Fields, A., Patterson, C. and Cyr, D.M. 2004. A foldable CFTR{Delta}F508 biogenic intermediate accumulates upon inhibition of the Hsc70-CHIP E3 ubiquitin ligase.*The Journal of Cell Biology*. 167(6): pp.1075-1085.

Yunus, A.A. and Lima, C.D. 2006. Lysine activation and functional analysis of E2-mediated conjugation in the SUMO pathway. *Nature Structural & Molecular Biology*. doi: 10.1038/nsmb1104.

Zhang, J. and Gu, X. 1998. Correlation between the substitution rate and rate variation among sites in protein evolution Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics*. 149(3): pp.1615-1625.

Zhang, X-D., Jenkins, J.N., Callahan, F.E., Creech, R.G., Yang, S., McCarty, J.C., Saha, S. and Ma, D-P., 2003. Molecular cloning, differential expression, and functional characterization of a family of class I ubiquitin-conjugating enzyme (E2) genes in cotton (Gossypium). *Biochemica Et Biophysica Acta*. 1625(3): pp.269-279.

Zmasek, C.M. and Eddy, S.R. 2001. ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*. 17(4): pp.383-384.

CHAPTER EIGHTEEN

APPENDIX

1) The accession numbers of all 193 peptide sequences of the phylogenetic trees.

Table 18.1
DATABASE ACCESSION NUMBERS OF THE ABBREVIATED NAMES GIVEN IN THE PHYLOGENETIC TREE:

NAMES	ACCESSION NO	NAMES	ACCESSION NO	NAMES	ACCESSION NO
atUBC1	Q9FI61	ceUBC8a_v1	AAF60891	mmHIP2	P27924
atUBC2a	P42745	ceUBC8a_v2	NP_500245	mmUB2D4	73318
atUBC2b_v1	P42746	ceUBC9	Q95017	MmUBE2B	P23567
atUBC2b_v2	Q8H1P1	ceUBC13	Q95XX0	MmUBE2R2	Q8VDE5
atUBC2b_v3	Q94AD2			MmUBE2D2	P51669
atUBC4a_v1	Q9SLE4	dmUBC1	P91633	MmUBE2J1	Q9DC92
atUBC4a_v2	P35132	dmUBC2	P25153	MmUBE2J2	Q9QX58
atUBC4b	P35131	dmUBC3a_v1	Q8IQM5	MmUBE2G2	Q9QX59
atUBC4c	P35132-2	dmUBC3a_v2	Q9VUR4	MmUBE2I	P50550
atUBC6a	Q9FK29	dmUBC4	P25867	MmUBE2a	Q9Z255
atUBC6b	Q9SHI7	dmUBC6	Q9VEJ4	MmUBE2C	Q9D1C1
atUBC6c	Q9LSP7	dmUBC7a	Q9VXE8	MmUBE2M	O76069
atUBC7a	P42747	dmUBC7b	Q8IH42	MmUBE2N	Q16781
atUBC7b	Q42540	dmUBC7c	Q9N9Z5	MmCDC34	Q8CFI2
atUBC7c	Q42540	dmUBC7d	Q8SYG3	MmUBC3	BAC35904
atUBC7d	AT5G59300	dmUBC8a	Q9W3K7	mmTSG101	Q61187
atUBC7e	Q42541	dmUBC8b	Q9VGD6	MmUBE2V2	Q8CE99
atUBC8a_v1	AAG52444	dmUBC9	O62622	MmUBE2S	Q921J4
atUBC8a_v2	P42749	dmUBC11	Q9VTY6	MmUBE2E3	P52483
atUBC8b	P42748	dmUBC12	Q9VSF3		
atUBC9	Q42551	dmUBC13a	P35128	ncUBC1	Q9P725
atUBC10	Q8LGF7	dmUBC13b	Q9VJS5	ncUBC2	P52493
atUBC11a_v1	Q9LPS2	dmUBC13c	Q9NKC1	ncUBC3a_v1	Q7S7Z9
atUBC11a_v2	Q9C6Q4			ncUBC3a_v2	Q8X0N3
atUBC11a_v3	Q8L7T3	hsHIP2	P27924	ncUBC6a	Q7S6X1
atUBC11b	Q9LJZ5	hsUBE2A	Q9Z255	ncUBC6b	Q7S7C4
atUBC12a	Q9SDY5	hsCDC34	P49427	ncUBC7	Q9C2A5
atUBC12b	Q9ZU75	hsUBE2R2	Q9NX64	ncUBC8	Q7S086
atUBC13a_v1	Q94A97	hsHBUCE1	Q9Y2X8	ncUBC9	Q871H3
atUBC13a_v2	Q9ZVA6	hsUBE2D2	P51669	ncUBC10	Q7SDB0
atUBC13b	Q9FZ48	hsUBE2D3	BAC04632	ncUBC12	Q8X0I8
		hsUBE2D1	P51668	ncUBC13	Q7SDM1
agUBC1	Q7PHN5	hsUBE2E1	P52483	osUBC1	Q8S1Y5
agUBC2a	Q7QFT1	hsUBE2J1	Q9NY66	osUBC2a_v1	Q7XIC8
agUBC2b	Q7PFQ9	hsUBE2J2_V1	Q96T84	osUBC2a_v2	NP_911360
agUBC4	Q7PFW9	hsUBE2J2_V2	Q96N26	OsUBC2b	Q8SBC1
agUBC6	Q7Q6A2	hsUBE2V2	Q15819	osUBC4	Q94DA8
agUBC7a	Q7QK63	hsUBE2V1	Q13403	osUBC7	Q943L1
agUBC7b	Q7PTL6	hsUBE2S	Q16763	osUBC8	Q7XDX9
agUBC8	Q7PJL8	hsUBE2E3	Q969T4	osUBC9a	Q7XCN3
agUBC9a	Q7QH85	hsUBE2E2	Q96LR5	osUBC9b	Q8H8G9
agUBC9b	Q7PZV4	hsUBE2G2_v1	AAP35560	osUBC9c	CAD41164
agUBC11a	Q7Q5S2	hsUBE2G2_v2	AAC32312	osUBC11	Q9LIY1

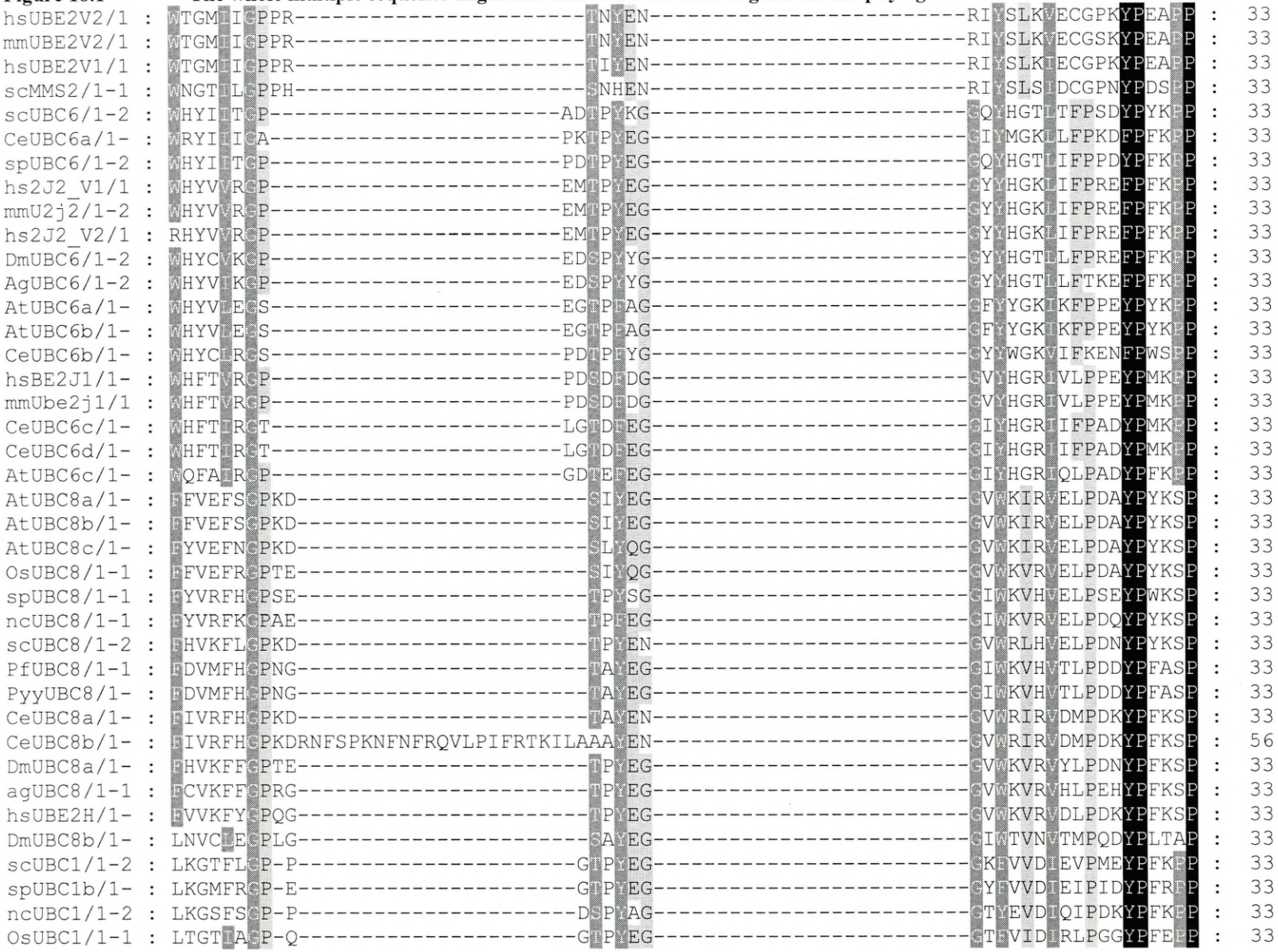
The accession numbers of all 193 peptide sequences of the phylogenetic trees continued.

Table 18.1
DATABASE ACCESSION NUMBERS OF THE ABBREVIATED NAMES GIVEN IN THE PHYLOGENETIC TREE:

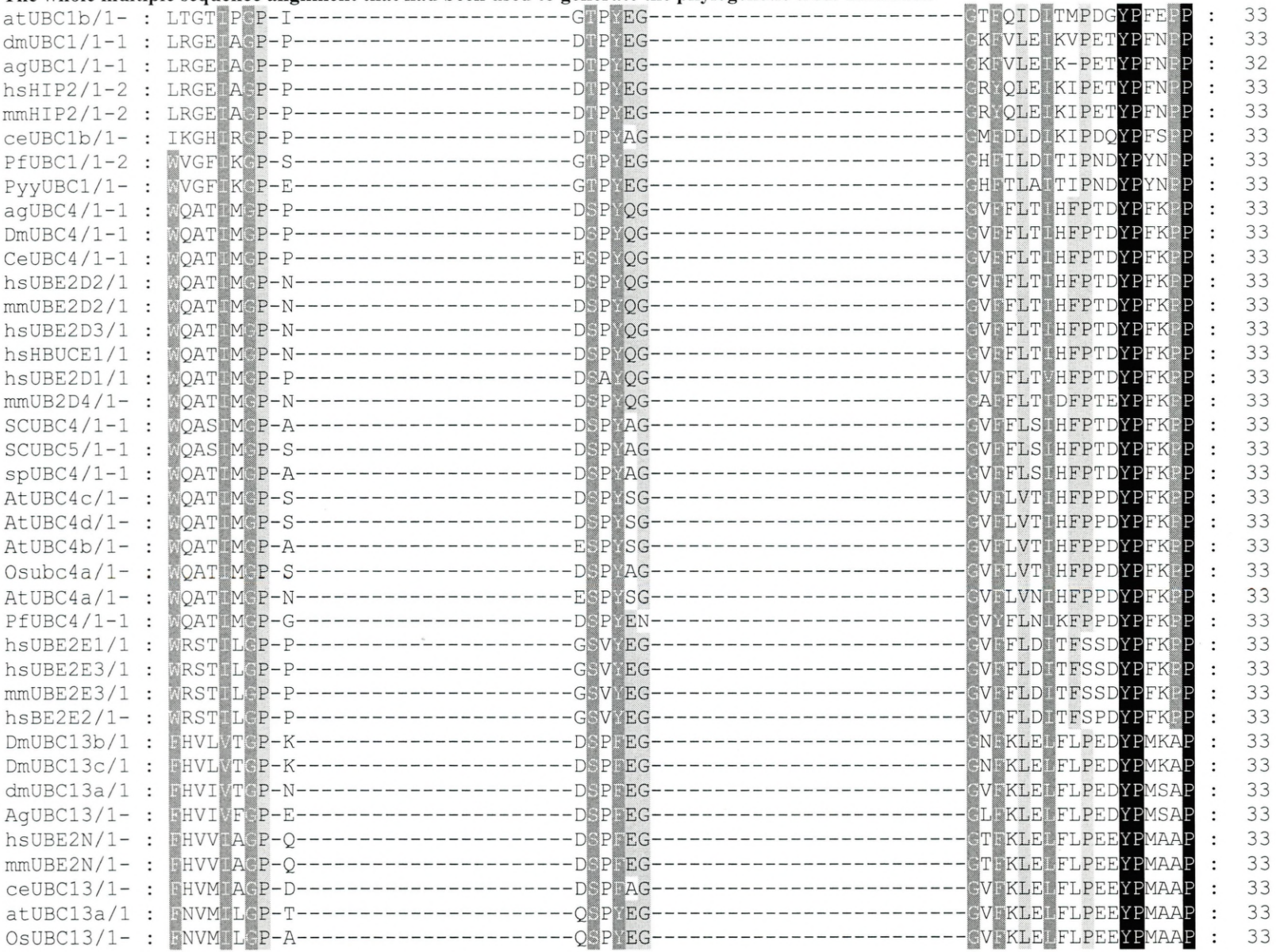
NAMES	ACCESSION NO.	NAMES	ACCESSION NO.	NAMES	ACC. NO.
agUBC12	Q7Q6C9	hsUBE2G1	Q99462	osUBC12	Q8LMK7
agUBC13	Q7Q9K0	hsUBE2L3	P51966	osUBC13	Q8W0I1
		hsUBE2I	P50550		
ceUBC1	NP_497174	hsQ9BQ25	Q9BQ25	pfUBC1	Q8IDD9
ceUBC2a_v1	Q9TZ69	hsUBE2B	P23567	pfUBC2a_v1	O77397
ceUBC2a_v2	P52478	hsUBE2C	O00762	pfUBC2a_v2	T18512
ceUBC3	Q95XN7	hsUBE2H	P37286	pfUBC2b	Q8IAW2
ceUBC4	P35129	hsUBE2L6	O14933	pfUBC4	Q8I607
ceUBC6a	NP_871922	hsUBE2M	O76069	pfUBC7	Q8I2U3
ceUBC6b	Q9N490	hsUBE2N	Q16781	pfUBC8	Q8IJ70
ceUBC6c_v1	T34195	hsTSG101	Q99816	pfUBC9	Q8I301
ceUBC6c_v2	NP_495566			pfUBC10	Q8ICV3
ceUBC7a	Q9U1Q1			pfUBC12	Q8I4X8
ceUBC7b	P34477			pfUBC13	Q8I3J4
scUBC1	P21734	scUBC9	P50623	pyyUBC1	Q7RP52
scUBC2	P06104	scUBC10	P29340	pyyUBC2a	Q7RNM4
scUBC3	P14682	scUBC11	P52492	pyyUBC2b	Q7RQW9
scUBC4	P15731	scUBC12	P52491	pyyUBC8	Q7RS94
scUBC5	P15732	scUBC13	P52490	pyyUBC10	Q7RJU9
scUBC6	P33296	scMMS2	P53152		
scUBC7	Q02159				
scUBC8	P28263				
spUBC1	O74810	spUBC8	Q9P7R4		
spUBC2	P23566	spUBC9	P40984		
spUBC4	P46595	spUBC10	Q9P6I1		
spUBC6	O42646	spUBC11	O00103		
spUBC7a	O00102	spUBC12	O74549		
spUBC7b	Q9Y818	spUBC13	O13685		

Table 18.1 illustrates all the peptide sequences of the phylogenetic tree, with their accession numbers. From the accession numbers, the peptide sequences can be retrieved from any of the databases like NCBI, EMBL and ENSEMBL.

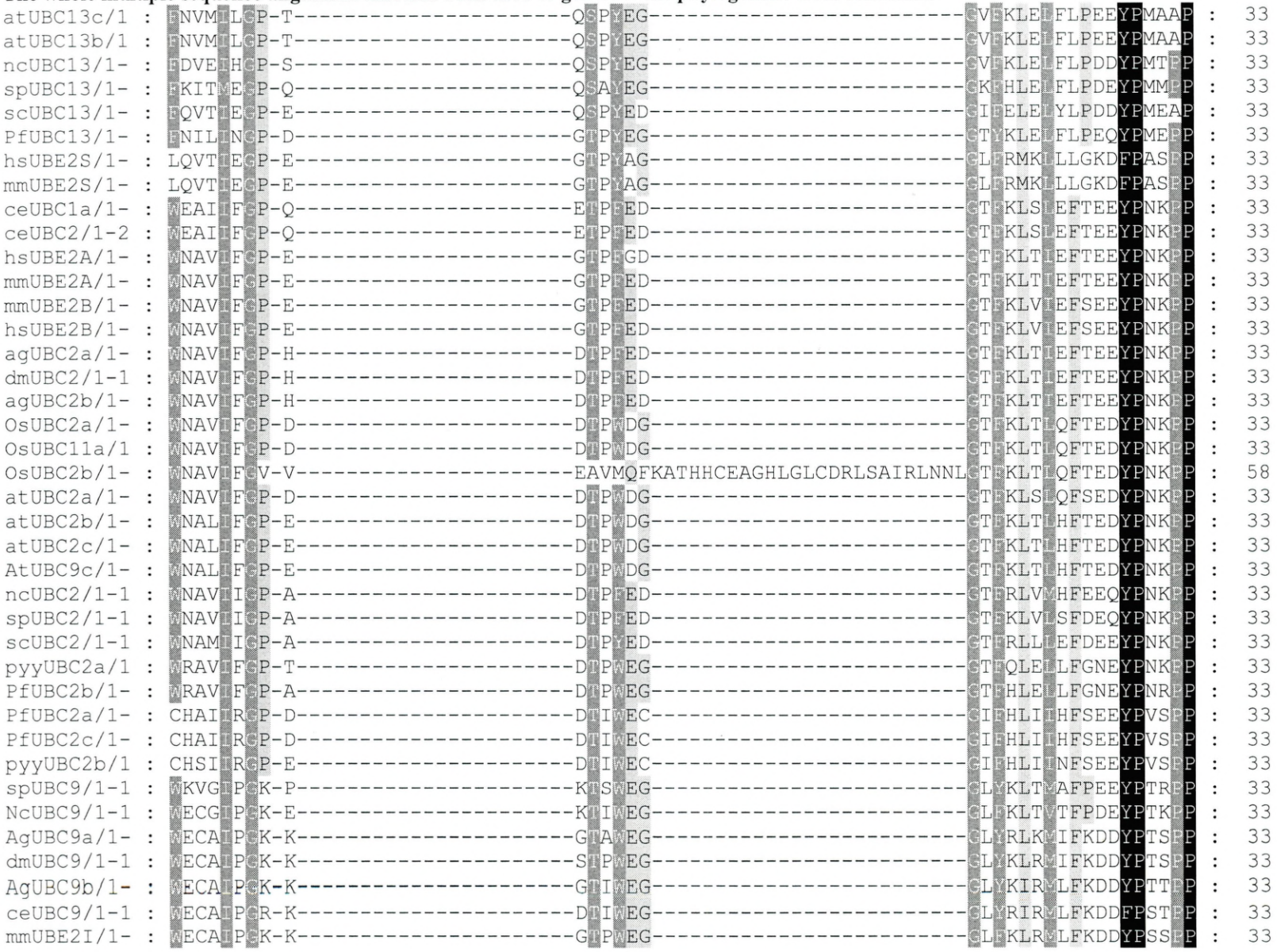
Figure 18.1 **The whole multiple sequence alignment that had been used to generate the phylogenetic trees:**



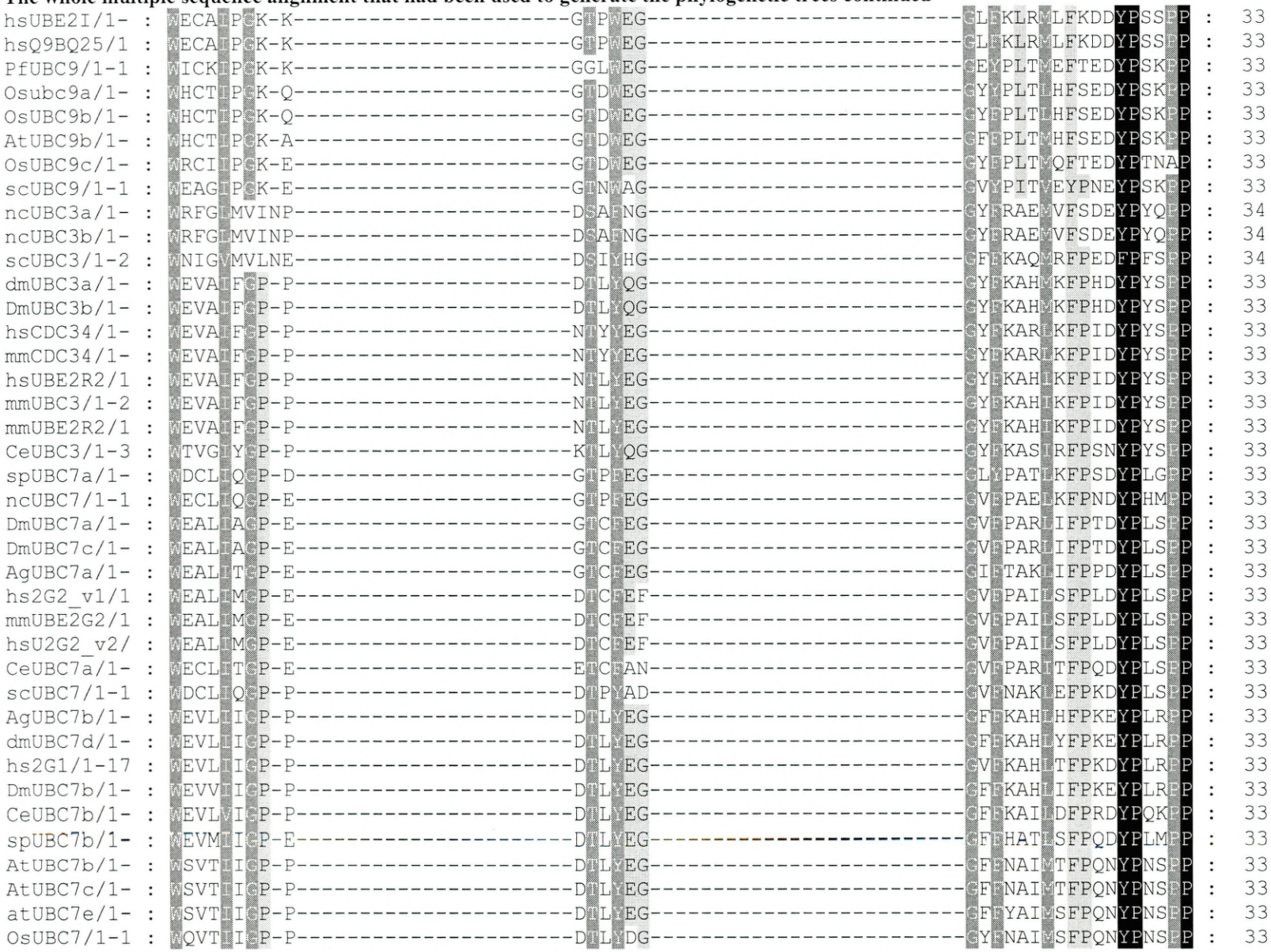
The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued



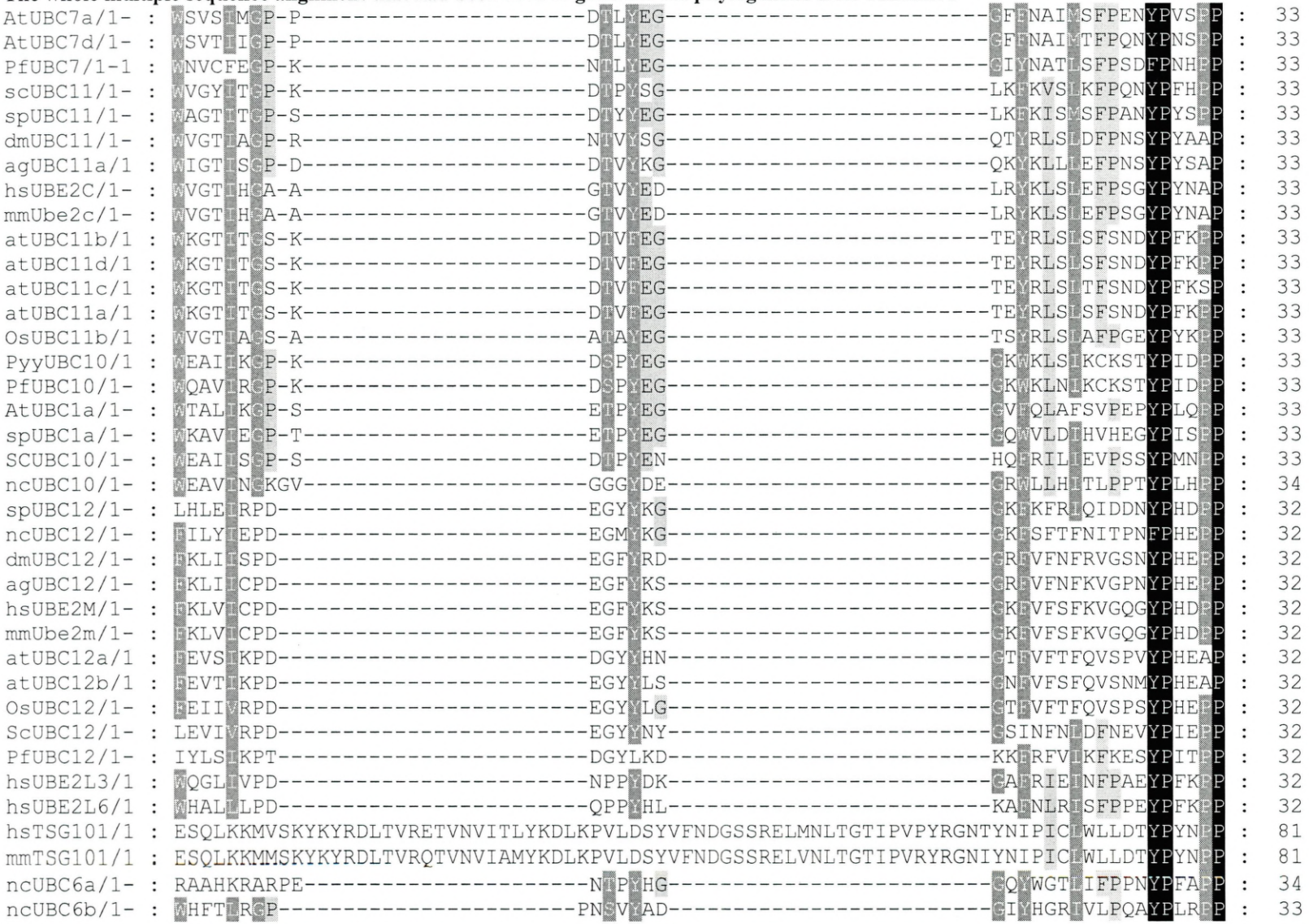
The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued



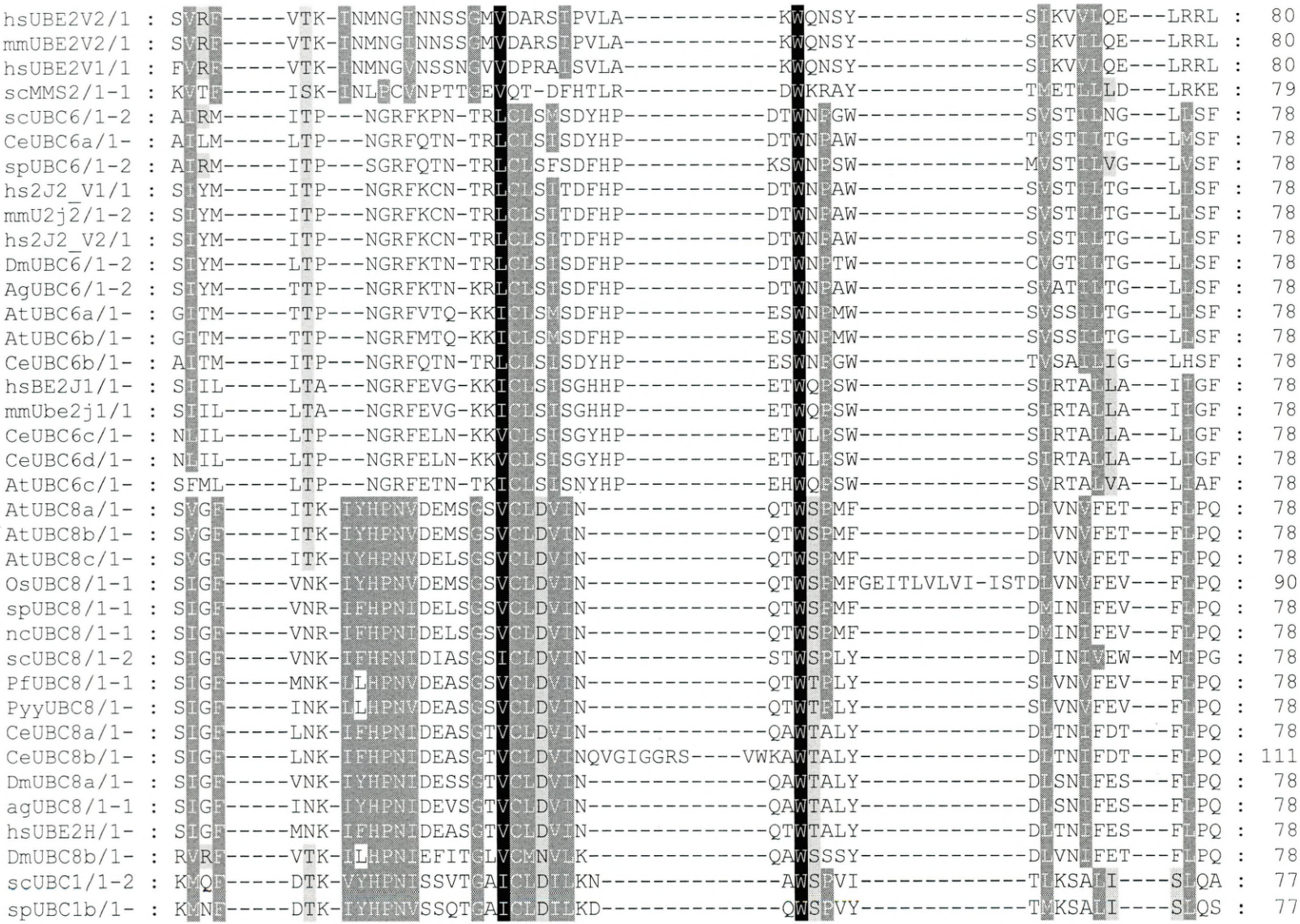
The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued



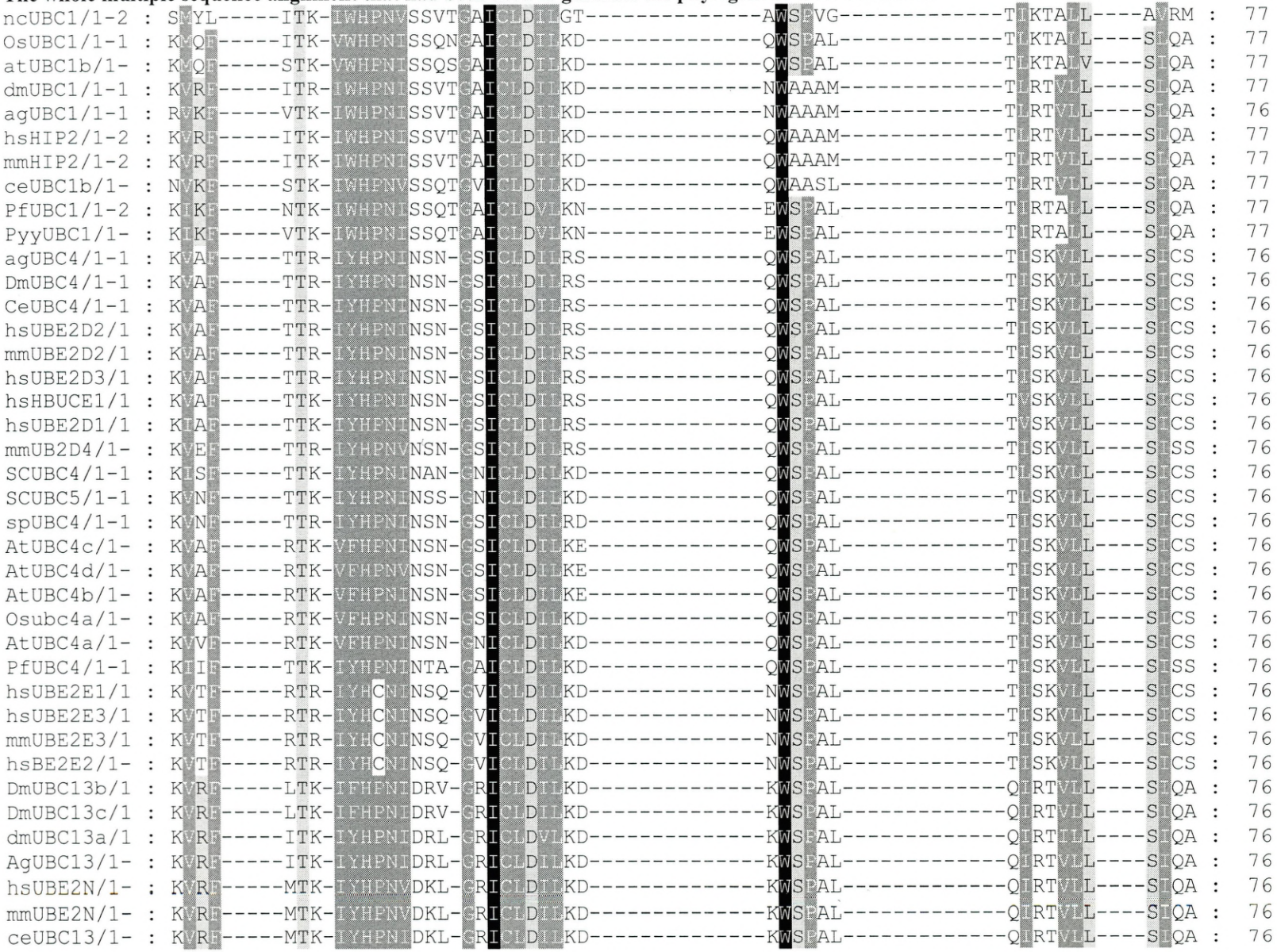
The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued



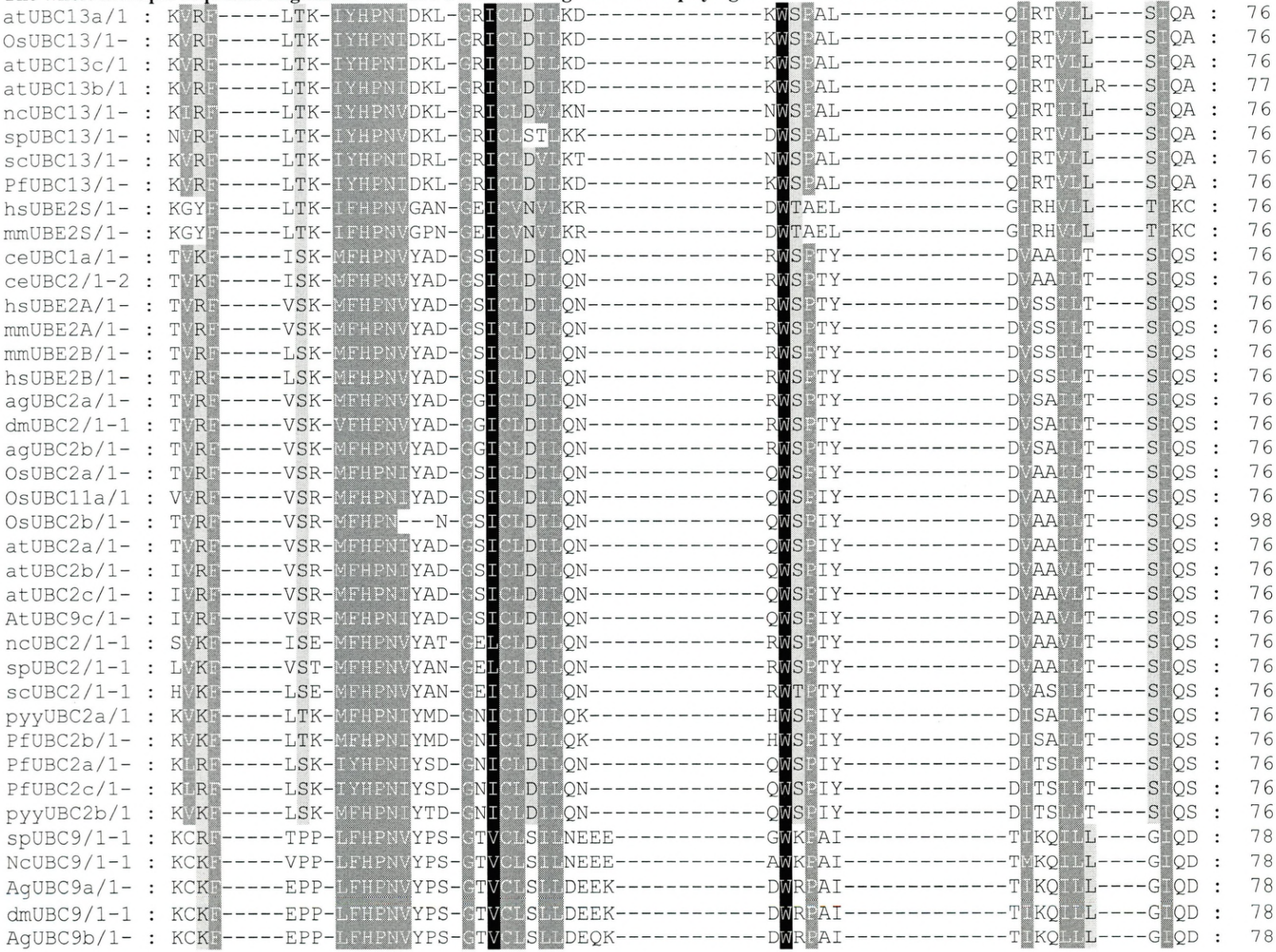
The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued



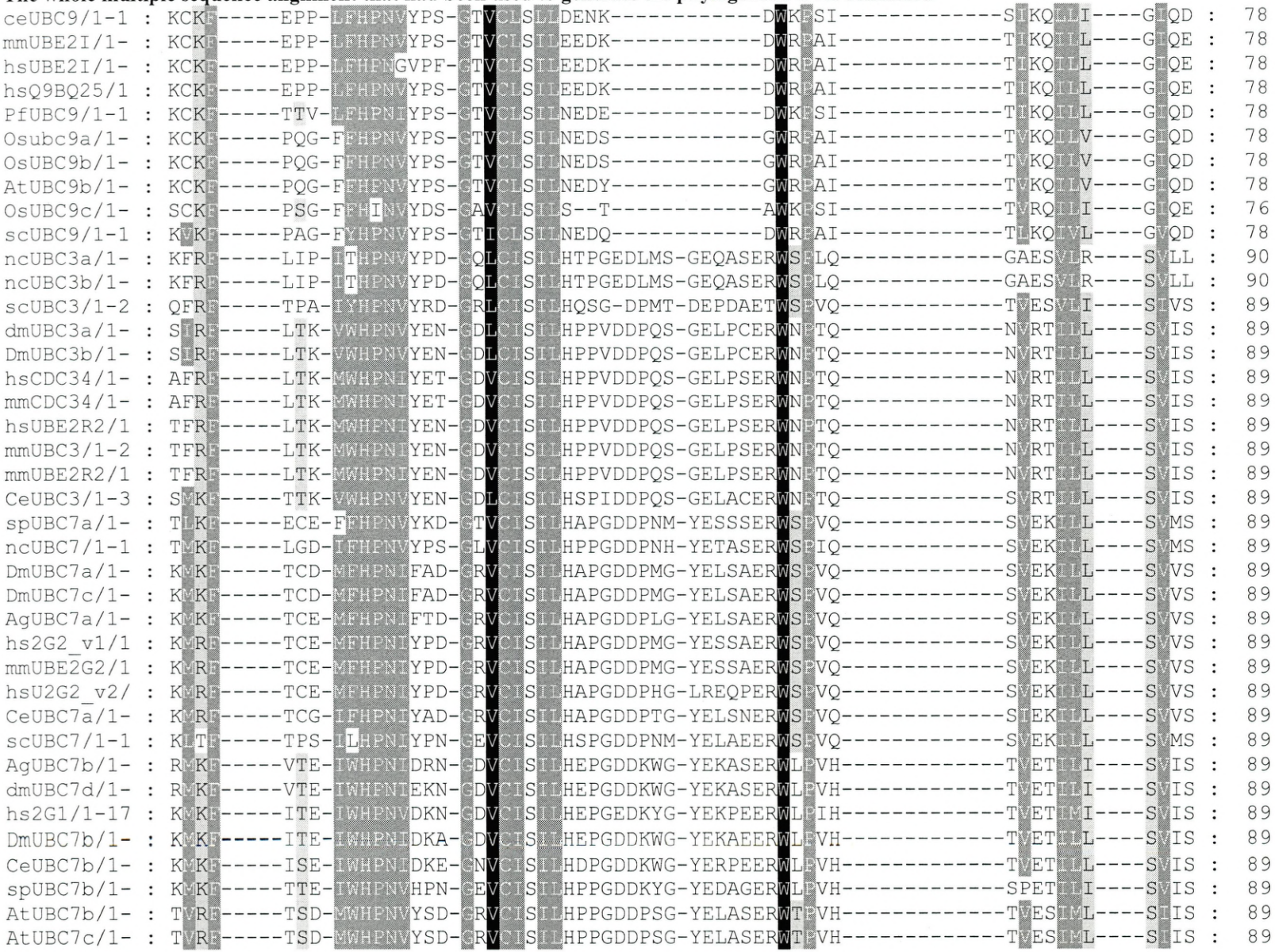
The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued



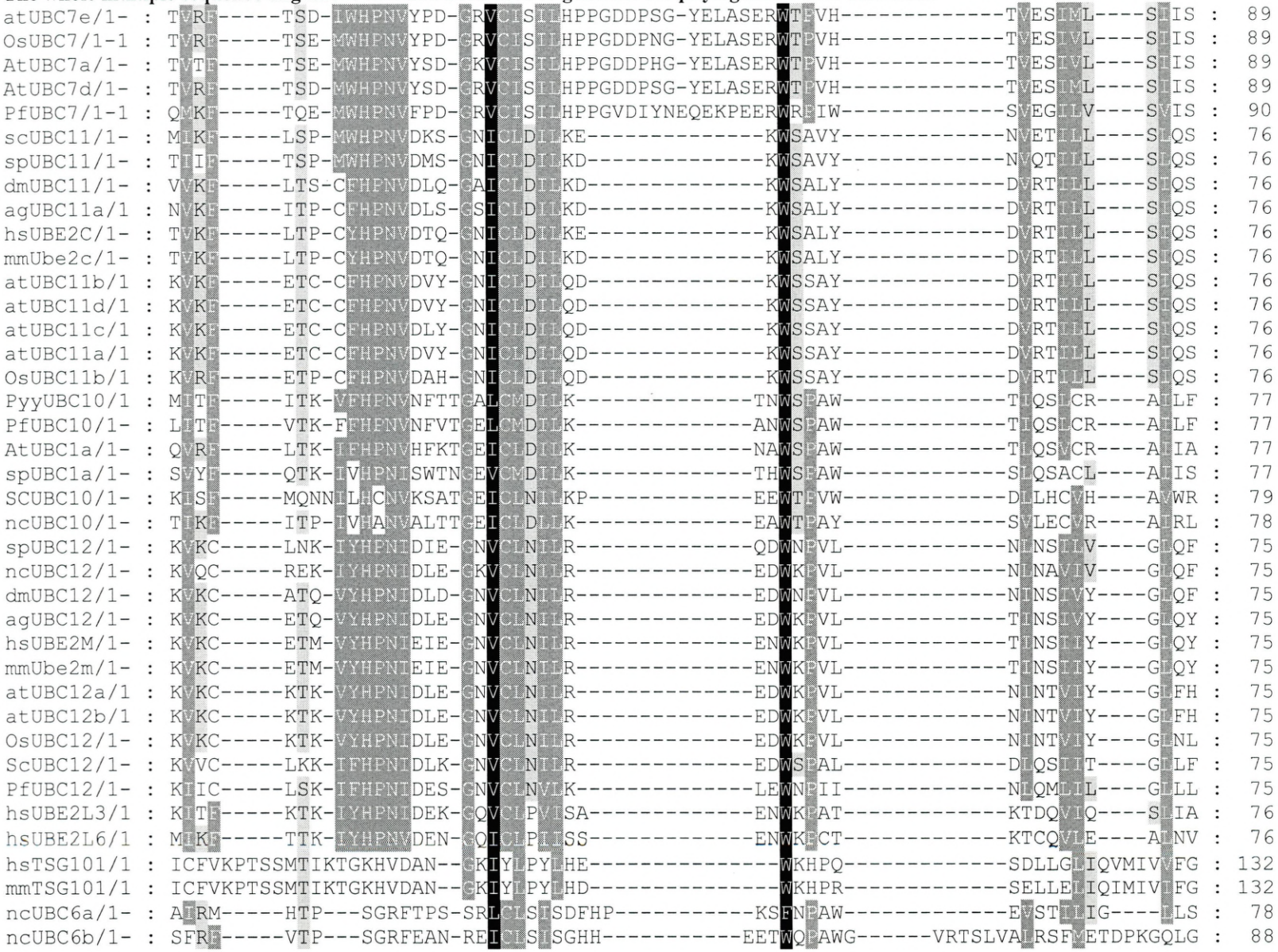
The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued



The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued



The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued



The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued

```
hsUBE2V2/1 : MMSKENMKLPQPPEGQTY :  98
mmUBE2V2/1 : MMSKENMKLPQPPEGQTY :  98
hsUBE2V1/1 : MMSKENMKLPQPPEGQCY :  98
scMMS2/1-1 : MATPANKKLRQPKEGETF :  97
scUBC6/1-2 : MTSD-EATTGSITTSDHQ :  95
CeUBC6a/1- : MNDN-QPTLGSLVTSESE :  95
spUBC6/1-2 : MTSD-EITTGGIVTSEST :  95
hs2J2_V1/1 : MVEK-GPTLGSIETSDFT :  95
mmU2j2/1-2 : MVEK-GPTLGSIETSDFT :  95
hs2J2_V2/1 : MVEK-GPTLGSIETSDFT :  95
DmUBC6/1-2 : MLES-TPTLGSIESSNYD :  95
AgUBC6/1-2 : MLES-TPTMGSCETTPAE :  95
AtUBC6a/1- : MMDN-SPTTGSVNTSVAE :  95
AtUBC6b/1- : MMDT-SPTTGSVNTTVIE :  95
CeUBC6b/1- : MNEN-SPAAGSIAGTPQD :  95
hsBE2J1/1- : MPTKGEGAIGSLDYTPEE :  96
mmUbe2j1/1 : MPTKGEGAIGSLDYTPEE :  96
CeUBC6c/1- : LPSTPGGALGSLDYPPKE :  96
CeUBC6d/1- : LPSTPGGALGSLDYPPKE :  96
AtUBC6c/1- : MPTSPNGALGSVDYPKDE :  96
AtUBC8a/1- : LLLYPNPSDPLNGEAAAL :  96
AtUBC8b/1- : LLLYPNPSDPLNGEAAAL :  96
AtUBC8c/1- : LLLYPNPSDPLNGEAAAL :  96
OsUBC8/1-1 : LLLYPNPSDPLNGEAAAL : 108
spUBC8/1-1 : LLRYPNASDPLNGEAAAL :  96
ncUBC8/1-1 : LLRYPNPTDPLNGEAAAM :  96
scUBC8/1-2 : LLKEPNGSDPLNNEAATL :  96
PfUBC8/1-1 : LLTYPNPSDPLNSDAASL :  96
PyyUBC8/1- : LLTYPNPSDPLNSDAASL :  96
CeUBC8a/1- : LLTYPNAADPLNGEAARL :  96
CeUBC8b/1- : LLTYPNAADPLNGEAARL : 129
DmUBC8a/1- : LLTYPNPVDPLNRDAAAL :  96
agUBC8/1-1 : LLTYPNPVDPLNGDAAAM :  96
hsUBE2H/1- : LLAYPNPIDPLNGDAAAM :  96
DmUBC8b/1- : LLRYPNPHDSLNHRAAAI :  96
scUBC1/1-2 : LLQSPEPNDPQDAEVAQH :  95
spUBC1b/1- : LLCTPEPSNPQDAQVAQV :  95
ncUBC1/1-2 : LLESPNPKDPQDAQVAKM :  95
```

The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued

```
OsUBC1/1-1 : LLSAPAPDDPQDAVVAQQ :  95
atUBC1b/1- : LLSAPEPKDPQDAVVAEQ :  95
dmUBC1/1-1 : LLAAAEPDDPQDAVVAYQ :  95
agUBC1/1-1 : LLAAAEPDDPQDAVVATQ :  94
hsHIP2/1-2 : LLAAAEPDDPQDAVVANQ :  95
mmHIP2/1-2 : LLAAAEPDDPQDAVVANQ :  95
ceUBC1b/1- : LMCTPEPKDPQDAVVAKQ :  95
PfUBC1/1-2 : LLSDPQPDDPQDAEVAKM :  95
PyyUBC1/1- : LLSDPQPDDPQDAEVAKM :  95
agUBC4/1-1 : LLCDPNPDDPLVPEIARI :  94
DmUBC4/1-1 : LLCDPNPDDPLVPEIARI :  94
CeUBC4/1-1 : LLCDPNPDDPLVPEIARI :  94
hsUBE2D2/1 : LLCDPNPDDPLVPEIARI :  94
mmUBE2D2/1 : LLCDPNPDDPLVPEIARI :  94
hsUBE2D3/1 : LLCDPNPDDPLVPEIARI :  94
hsHBUCE1/1 : LLCDPNPDDPLVPEIAHT :  94
hsUBE2D1/1 : LLCDPNPDDPLVPDIAQI :  94
mmUB2D4/1- : LLCDPNPDDPLVPEIAQI :  94
SCUBC4/1-1 : LLTDANPDDPLVPEIAHI :  94
SCUBC5/1-1 : LLTDANPDDPLVPEIAQI :  94
spUBC4/1-1 : LLTDPNPDDPLVPEIAHV :  94
AtUBC4c/1- : LLTDPNPDDPLVPEIAHM :  94
AtUBC4d/1- : LLTDPNPDDPLVPEIAHM :  94
AtUBC4b/1- : LLTDPNPDDPLVPEIAHM :  94
Osubc4a/1- : LLTDPNPDDPLVPEIAHM :  94
AtUBC4a/1- : LLTDPNPDDPLVPEIAHI :  94
PfUBC4/1-1 : LLTDPNADDPLVPEIAHV :  94
hsUBE2E1/1 : LLTDCNPADPLVGSIATQ :  94
hsUBE2E3/1 : LLTDCNPADPLVGSIATQ :  94
mmUBE2E3/1 : LLTDCNPADPLVGSIATQ :  94
hsBE2E2/1- : LLTDCNPADPLVGSIATQ :  94
DmUBC13b/1 : LLSAPNPDDPLANDVAEL :  94
DmUBC13c/1 : LLSAPNPDDPLANDVAEL :  94
dmUBC13a/1 : LLSAPNPDDPLANDVAEL :  94
AgUBC13/1- : LLSAPNPDDPLANDVAEL :  94
hsUBE2N/1- : LLSAPNPDDPLANDVAEQ :  94
mmUBE2N/1- : LLSAPNPDDPLANDVAEQ :  94
ceUBC13/1- : LLSAPNPEDPLATDVAEQ :  94
atUBC13a/1 : LLSAPNPDDPLSENIAKH :  94
```

The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued

```
OsUBC13/1- : LLSAPNPDDPLSDNIAKH :  94
atUBC13c/1 : LLSAPNPDDPLSENIAKH :  94
atUBC13b/1 : LLSAPNPDDPLSENIAKH :  95
ncUBC13/1- : LLGAPNPDDPLAPDVAKT :  94
spUBC13/1- : LMGAPNPDDPLDNDVAKI :  94
scUBC13/1- : LLASPNPNDPLANDVAED :  94
PfUBC13/1- : LLSSPEPDDPLDSKVAEH :  94
hsUBE2S/1- : LLIHPNPESALNEEAGRL :  94
mmUBE2S/1- : LLIHPNPESALNEEAGRL :  94
ceUBC1a/1- : LLDEPNPNSPANSLAAQL :  94
ceUBC2/1-2 : LLDEPNPNSPANSLAAQL :  94
hsUBE2A/1- : LLDEPNPNSPANSQAAQL :  94
mmUBE2A/1- : LLDEPNPNSPANSQAAQL :  94
mmUBE2B/1- : LLDEPNPNSPANSQAAQL :  94
hsUBE2B/1- : LLDEPNPNSPANSQAAQL :  94
agUBC2a/1- : LLSDPNPNSPANSMAAQL :  94
dmUBC2/1-1 : LLSDPNPNSPANSTAAQL :  94
agUBC2b/1- : LLSDPNPNSPANSMAAQL :  94
OsUBC2a/1- : LLCDPNPNSPANSEAARL :  94
OsUBC11a/1 : LLCDPNPNSPANSEAARL :  94
OsUBC2b/1- : LLCDPNPNSPANSEAARL : 116
atUBC2a/1- : LLCDPNPNSPANSEAARM :  94
atUBC2b/1- : LLCDPNPDSPANAEAARL :  94
atUBC2c/1- : LLCDPNPDSPANAEAARL :  94
AtUBC9c/1- : LLCDPNPDSPANAEAARL :  94
ncUBC2/1-1 : LLNDPNTGSPANVEASNL :  94
spUBC2/1-1 : LLNDPNNASPANAEAAQL :  94
scUBC2/1-1 : LFNDPNPASPANVEAATL :  94
pyyUBC2a/1 : LLSDPNPNSPANQEAALL :  94
PfUBC2b/1- : LLSDPNPNSPANQEAALL :  94
PfUBC2a/1- : LLNDPNTSSPANPEAARI :  94
PfUBC2c/1- : LLNDPNTSSPANPEAARI :  94
pyyUBC2b/1 : LLNDPNTASPANPEAAKI :  94
spUBC9/1-1 : LLDDPNIASPAQTEAYTM :  96
NcUBC9/1-1 : LLNDPNPESPAQAEAYNL :  96
AgUBC9a/1- : LLNEPNIKDPAQAEAYTI :  96
dmUBC9/1-1 : LLNEPNIKDPAQAEAYTI :  96
AgUBC9b/1- : LLNEPNIKDPAQAEAYTI :  96
ceUBC9/1-1 : LLNHPNIEDPAQAEAYQI :  96
```

The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued

```
mmUBE2I/1-  :  LLNEPNIQDPAQAEAYTI  :  96
hsUBE2I/1-  :  LLNEPNIQDPAQAEAYTI  :  96
hsQ9BQ25/1  :  LLNEPNIQDPAQAEAYTI  :  96
PfUBC9/1-1  :  LLDNPNPNSPAQAEPFLL  :  96
Osubc9a/1-  :  LLDQPNPADPAQTDGYHI  :  96
OsUBC9b/1-  :  LLDQPNPADPAQTDGYHL  :  96
AtUBC9b/1-  :  LLDTPNPADPAQTDGYHL  :  96
OsUBC9c/1-  :  LFDDPNPNSAAQNISYEL  :  94
scUBC9/1-1  :  LLDSPNPNSPAQEPAWRS  :  96
ncUBC3a/1-  :  LLDDPEINSPANVDAGVM  : 108
ncUBC3b/1-  :  LLDDPEINSPANVDAGVM  : 108
scUBC3/1-2  :  LLEDPNINSPANVDAAVD  : 107
dmUBC3a/1-  :  LLNEPNTFSPANVDASVM  : 107
DmUBC3b/1-  :  LLNEPNTFSPANVDASVM  : 107
hsCDC34/1-  :  LLNEPNTFSPANVDASVM  : 107
mmCDC34/1-  :  LLNEPNTFSPANVDASVM  : 107
hsUBE2R2/1  :  LLNEPNTFSPANVDASVM  : 107
mmUBC3/1-2  :  LLNEPNTFSPANVDASVM  : 107
mmUBE2R2/1  :  LLNEPNTFSPANVDASVM  : 107
CeUBC3/1-3  :  LLNEPNTSSPANVDASVM  : 107
spUBC7a/1-  :  MLAEPNDESGANIDACKM  : 107
ncUBC7/1-1  :  MLAEPNDESPANVEAAKM  : 107
DmUBC7a/1-  :  MLAEPNDESGANVDAAIM  : 107
DmUBC7c/1-  :  MLAEPNDESGANVDAAIM  : 107
AgUBC7a/1-  :  MLAEPNDESGANVDAAIM  : 107
hs2G2_v1/1  :  MLAEPNDESGANVDASKM  : 107
mmUBE2G2/1  :  MLAEPNDESGANVDASKM  : 107
hsU2G2_v2/  :  MLAEPNDESGANVDASKM  : 107
CeUBC7a/1-  :  MLAEPNDESPANVSAAKM  : 107
scUBC7/1-1  :  MLSEPNIESGANIDACIL  : 107
AgUBC7b/1-  :  MLADPNDESPANVDAAKE  : 107
dmUBC7d/1-  :  MLADPNDESPANVDAAKE  : 107
hs2G1/1-17  :  MLADPNGDSPANVDAAKE  : 107
DmUBC7b/1-  :  MLTDPNDESAANVDAAKE  : 107
CeUBC7b/1-  :  MLTDPNFESPANVDAAKM  : 107
spUBC7b/1-  :  MLSSPNDESPANIDAAKE  : 107
AtUBC7b/1-  :  MLSGPNDESPANVEAAKE  : 107
AtUBC7c/1-  :  MLSGPNDESPANVEAAKE  : 107
atUBC7e/1-  :  MLSGPNDESPANVEAAKE  : 107
```

The whole multiple sequence alignment that had been used to generate the phylogenetic trees continued

```
OsUBC7/1-1 : MLSGPNDESPANIEAAKE : 107
AtUBC7a/1- : MLSGPNDESPANVEAAKE : 107
AtUBC7d/1- : MLSGPNDESPANVEAAKE : 107
PfUBC7/1-1 : MLNEPNLESPANVDAAVQ : 108
scUBC11/1- : LLGEPNNRSPLNAVAAEL :  94
spUBC11/1- : LLGEPNNASPLNAQAAEL :  94
dmUBC11/1- : LLGEPNNESPLNAQAAMM :  94
agUBC11a/1 : LLGEPNNDSPLNSQASQL :  94
hsUBE2C/1- : LLGEPNIDSPLNTHAAEL :  94
mmUbe2c/1- : LLGEPNIDSPLNTHAAEL :  94
atUBC11b/1 : LLGEPNISSPLNTQAAQL :  94
atUBC11d/1 : LLGEPNISSPLNTQAAQL :  94
atUBC11c/1 : LLGEPNISSPLNNQAAQL :  94
atUBC11a/1 : LLGEPNISSPLNTQAAQL :  94
OsUBC11b/1 : LLGEPNNDSPLNTQAAAL :  94
PyyUBC10/1 : LLNEPNAESPLNCDAGNL :  95
PfUBC10/1- : LFNEPNADSPLNCDAGNL :  95
AtUBC1a/1- : LMAHPEPDSPLNCDSGNL :  95
spUBC1a/1- : LLSNYDASSPLNVDAAKL :  95
SCUBC10/1- : LLREPVCDSPLDVDIGNI :  97
ncUBC10/1- : LLSCPGIDSPLNVDVAAL :  96
spUBC12/1- : LFLSPNAEDPLNKEAAAD :  93
ncUBC12/1- : LFLEPNASDPLNKEAAED :  93
dmUBC12/1- : LFLEPNPEDPLNKEAADV :  93
agUBC12/1- : LFLEPNPEDPLNREAAEV :  93
hsUBE2M/1- : LYLEPNPEDPLNKKAAEV :  93
mmUbe2m/1- : LYLEPNPEDPLNKKAAEV :  93
atUBC12a/1 : LFTEPNSEDPLNHDAAAV :  93
atUBC12b/1 : LFTEPNYEDPLNHEAAAV :  93
OsUBC12/1- : LFTQPNDEDPLNHEAAAV :  93
ScUBC12/1- : LFLEPNPNDPLNKDAAKL :  93
PfUBC12/1- : LIDEPSTDDPFNKIAAEV :  93
hsUBE2L3/1 : LVNDPQPEHPLRADLAEE :  94
hsUBE2L6/1 : LVNRPNIREPLRMDLADL :  94
hsTSG101/1 : DEPPVFSRPISASYPPYQ : 150
mmTSG101/1 : EEPPVFSRPTVSASYPPY : 150
ncUBC6a/1- : FMTSEEMTTGSVSATETE :  96
ncUBC6b/1- : GLDATEAVRRRMATESRA : 106
```

CRYSTALLOGRAPHY

All the crystallization conditions used: (Salts and Buffers); its corresponding precipitants are illustrated in the later pages as it was not possible to fit in all together.

Table 18.2

Hampton Research Crystal Screen formulation - Copyright 2001

Reagent number	[Salt]	[Salt] units	Salt	[Buffer]	[Buffer] units	Buffer	pH
1	0.02	M	calcium chloride dihydrate	0.1	M	sodium acetate trihydrate	4.6
2							
3							
4				0.1	M	Tris hydrochloride	8.5
5	0.2	M	sodium citrate tribasic dihydrate	0.1	M	Hepes sodium	7.5
6	0.2	M	magnesium chloride hexahydrate	0.1	M	Tris hydrochloride	8.5
7				0.1	M	sodium cacodylate	6.5
8	0.2	M	sodium citrate tribasic dihydrate	0.1	M	sodium cacodylate	6.5
9	0.2	M	ammonium acetate	0.1	M	sodium citrate tribasic dihydrate	5.6
10	0.2	M	ammonium acetate	0.1	M	sodium acetate trihydrate	4.6
11				0.1	M	sodium citrate tribasic dihydrate	5.6
12	0.2	M	magnesium chloride hexahydrate	0.1	M	Hepes sodium	7.5
13	0.2	M	sodium citrate tribasic dihydrate	0.1	M	Tris hydrochloride	8.5
14	0.2	M	calcium chloride dihydrate	0.1	M	Hepes sodium	7.5
15	0.2	M	ammonium sulfate	0.1	M	sodium cacodylate	6.5
16				0.1	M	Hepes sodium	7.5
17	0.2	M	lithium sulfate monohydrate	0.1	M	Tris hydrochloride	8.5
18	0.2	M	magnesium acetate tetrahydrate	0.1	M	sodium cacodylate	6.5

crystallization conditions continued

Reagent no.	Salts	salt units	salts	Buffer	Buffer units	Buffer	pH
19	0.2	M	ammonium acetate	0.1	M	Tris hydrochloride	8.5
20	0.2	M	ammonium sulfate	0.1	M	Sodium acetate trihydrate	4.6
21	0.2	M	magnesium acetate tetrahydrate	0.1	M	sodium cacodylate	6.5
22	0.2	M	sodium acetate trihydrate	0.1	M	Tris hydrochloride	8.5
23	0.2	M	magnesium chloride hexahydrate	0.1	M	Hepes sodium	7.5
24	0.2	M	calcium chloride dihydrate	0.1	M	Sodium acetate trihydrate	4.6
25				0.1	M	imidazole	6.5
26	0.2	M	ammonium acetate	0.1	M	sodium citrate tribasic dihydrate	5.6
27	0.2	M	sodium citrate tribasic dihydrate	0.1	M	Hepes sodium	7.5
28	0.2	M	sodium acetate trihydrate	0.1	M	sodium cacodylate	6.5
29				0.1	M	Hepes sodium	7.5
30	0.2	M	ammonium sulfate				
31	0.2	M	ammonium sulfate				
32							
33							
34				0.1	M	sodium acetate trihydrate	4.6
35				0.1	M	Hepes sodium	7.5
36				0.1	M	Tris hydrochloride	8.5
37				0.1	M	sodium acetate trihydrate	4.6
38				0.1	M	Hepes sodium	7.5
39				0.1	M	Hepes sodium	7.5
40				0.1	M	sodium citrate tribasic dihydrate	5.6
41				0.1	M	Hepes sodium	7.5
42	0.05	M	potassium phosphate monobasic				
43							
44							

crystallization conditions continued:

Reagent no.	Salts	salt units	salts	Buffer	Buffer units	Buffer	pH
45	0.2	M	zinc acetate dihydrate	0.1	M	sodium cacodylate	6.5
46	0.2	M	calcium acetate hydrate	0.1	M	sodium cacodylate	6.5
47				0.1	M	sodium acetate trihydrate	4.6
48				0.1	M	Tris hydrochloride	8.5
49	1	M	lithium sulfate monohydrate				
50	0.5	M	lithium sulfate monohydrate				

crystallization conditions used: (Precipitants)

Reagent no.	[Precipitant 1]	[Precipitant 1] units	Precipitant 1	[Ppt 2]	[Ppt 2] units	Ppt 2
1	30	%v/v	2-methyl-2,4-pentanediol			
2	0.4	M	potassium sodium tartrate tetrahydrate			
3	0.4	M	ammonium phosphate monobasic			
4	2	M	ammonium sulfate			
5	30	%v/v	2-methyl-2,4-pentanediol			
6	30	%w/v	polyethylene glycol 4000			
7	1.4	M	sodium acetate trihydrate			
8	30	%v/v	iso-propanol			
9	30	%w/v	polyethylene glycol 4000			
10	30	%w/v	polyethylene glycol 4000			
11	1	M	ammonium phosphate monobasic			
12	30	%v/v	iso-propanol			
13	30	%v/v	polyethylene glycol 400			
14	28	%v/v	polyethylene glycol 400			
15	30	%w/v	polyethylene glycol 8000			
16	1.5	M	lithium sulfate monohydrate			
17	30	%w/v	polyethylene glycol 4000			

crystallization conditions used: (Precipitants)						
Reagent #	[Precipitant 1]	[Precipitant 1] units	Precipitant 1	[Precipitant 2]	[Ppt 2] units	Precipitant 2
18	20	%w/v	polyethylene glycol 8000			
19	30	%v/v	iso-propanol			
20	25	%w/v	polyethylene glycol 4000			
21	30	%v/v	2-methyl-2,4-pentanediol			
22	30	%w/v	polyethylene glycol 4000			
23	30	%v/v	polyethylene glycol 400			
24	20	%v/v	iso-propanol			
25	1	M	sodium acetate trihydrate			
26	30	%v/v	2-methyl-2,4-pentanediol			
27	20	%v/v	iso-propanol			
28	30	%w/v	polyethylene glycol 8000			
29	0.8	M				
30	30	%w/v	polyethylene glycol 8000			
31	30	%w/v	polyethylene glycol 4000			
32	2	M	ammonium sulfate			
33	4	M	sodium formate			
34	2	M	sodium formate			
35	0.8	M	sodium phosphate monobasic monohydrate	0.8	M	Potassium phosphate monobasic
36	8	%w/v	polyethylene glycol 8000			
37	8	%w/v	polyethylene glycol 4000			
38	1.4	M	sodium citrate tribasic dihydrate			
39	2	%v/v	polyethylene glycol 400	2	M	Ammonium sulfate
40	20	%v/v	iso-propanol	20	%w/v	Polyethylene glycol 4000
41	10	%v/v	iso-propanol	20	%w/v	Polyethylene glycol 4000
42	20	%w/v	polyethylene glycol 8000			
43	30	%w/v	polyethylene glycol 1500			